

https://dspace.org/

The distalproximal relationships among the human moonlighting proteins Evolutionary hotspots and Darwinian checkpoints

Debaleena Nawn; Sk. Sarif Hassan; Moumita Sil; Ankita Ghosh; Pallab Basu ...

ELSEVIER SCIENCE BV
https://hdl.handle.net/10539/40708
WOS

Downloaded from a DSpace repository. https://dspace.org/

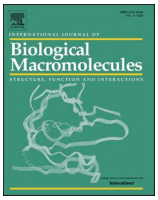
International Journal of Biological Macromolecules 259 (2024) 128998



Contents lists available at ScienceDirect

International Journal of Biological Macromolecules

journal homepage: www.elsevier.com/locate/ijbiomac





The distal-proximal relationships among the human moonlighting proteins: Evolutionary hotspots and Darwinian checkpoints

Debaleena Nawn [a,h,1], Sk. Sarif Hassan [b,1], Moumita Sil [a], Ankita Ghosh [a], Arunava Goswami [a], Pallab Basu [c,d], Guy W. Dayhoff II [e], Kenneth Lundstrom [f], Vladimir N. Uversky [g,*]

[a] Biological Science Division, Indian Statistical Institute, 203 B.T Road, Kolkata, 700108, West Bengal, India
[b] Department of Mathematics, Pingla Thana Mahavidyalaya, Maligram, Paschim Medinipur, West Bengal, India
[c] School of Physics, University of the Witwatersrand, Johannesburg, Braamfontein 2000, South Africa
[d] Woxsen School of Sciences, Woxsen University, Hyderabad 500 033, Telangana, India
[e] Department of Chemistry, University of South Florida, Tampa, FL 33612, USA
[f] PanTherapeutics, Rte de Lavaux 49, CH1095 Lutry, Switzerland
[g] Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA
[h] Indian Research Institute for Integrated Medicine (IRIIM), Unsani, Howrah, 711302, West Bengal, India

ARTICLE INFO

Keywords:
Moonlighting proteins
Phylogeny
Distal-proximal
Promiscuous proteins
Signature-features

ABSTRACT

Moonlighting proteins, known for their ability to perform multiple, often unrelated functions within a single polypeptide chain, challenge the traditional "one gene, one protein, one function" paradigm. As organisms evolved, their genomes remained relatively stable in size, but the introduction of post-translational modifications and sub-strategies like protein promiscuity and intrinsic disorder enabled multifunctionality. Enzymes, in particular, exemplify this phenomenon, engaging in unrelated processes alongside their primary catalytic roles. This study employs a systematic, quantitative informatics approach to shed light on human moonlighting protein sequences. Phylogenetic analyses of human moonlighting proteins are presented, elucidating the distal-proximal relationships among these proteins based on sequence-derived quantitative features. The findings unveil the captivating world of human moonlighting proteins, urging further investigations in the emerging field of moonlighting proteomics, with the potential for significant contributions to our understanding of multifunctional proteins and their roles in diverse cellular processes and diseases.

1. Introduction

In English, the term "moonlighting" refers to having an additional job alongside the primary task. In scientific terminology, the concept of moonlighting extends to proteins, where a protein, in addition to its main function, performs an extra function. Moonlighting proteins are unique in that they fulfill multiple independent and often unrelated tasks without segregating these functions into distinct protein domains [1,2]. As a result, proteins that emerge from gene fusions and have various functions are disregarded. Proteins that are different splice variants from the same gene are also excluded. Also, both functions should be independent, any changes in one function due to mutation should not affect the other functionality and vice versa. To delineate the phenomenon of moonlighting, Piatigorsky initially coined the term "gene sharing" [3].The term "moonlighting proteins" was first used by Jeffery [4,5]. The first ever moonlighting function was reported by Piatigorsky and Wistow in 1980s when they reported that certain crystallins, structural lens protein in the vertebrate eye had other enzymatic functions [6]. In recent years, the conventional paradigm of "one gene encoding one protein with a single function" has yielded to a more nuanced understanding that many proteins exhibit multifunctionality [7]. Among these multifunctional proteins, the category of moonlighting proteins stands as a captivating and intricate example. Moonlighting proteins are notable for their capacity to perform two disparate functions within a solitary polypeptide chain. The hypothesis by Beadle and Tatum stating one gene-one polypeptide-one function cannot be applied

* Corresponding author.
E-mail addresses: debaleena.nawn@gmail.com (D. Nawn), sksarifhassan@pinglacollege.ac.in (Sk.S. Hassan), moumitasil20@gmail.com (M. Sil), ankitaghosh100@gmail.com (A. Ghosh), srabanisopanarunava@gmail.com (A. Goswami), pallabbasu@gmail.com (P. Basu), gdayhoff@usf.edu (G.W. Dayhoff), lundstromkenneth@gmail.com (K. Lundstrom), vuversky@usf.edu (V.N. Uversky).
[1] These authors contributed equally in the work.

https://doi.org/10.1016/j.ijbiomac.2023.128998
Received 14 October 2023; Received in revised form 19 December 2023; Accepted 21 December 2023
Available online 3 January 2024
0141-8130/© 2024 Elsevier B.V. All rights reserved.

Table 1
List of Moonlighting proteins in humans with their associated Uniprot ID (hyperlinked with respective Uniprot webpage).

Serial No.	Uniprot ID	Moonlighting proteins of human
1	P28482	ERK2 (signal-regulated kinases)
2	Q99714	3-hydroxyacyl-CoA dehydrogenase type-2
3	P55263	Adenosine kinase
4	Q7L0Y3	Mitochondrial RNase P
5	P47897	Glutamine–tRNA ligase
6	Q02252	Methylmalonate-semialdehyde dehydrogenase [acylating], mitochondrial
7	P19971	Thymidine phosphorylase
8	P06744	Phosphoglucose isomerase
9	P62826	RanGTP (GTP-binding nuclear protein Ran)
10	P32189	Glycerol kinase
11	P78545	ETS-related transcription factor Elf-3
12	Q15046	Lysine–tRNA ligase
13	P21912	Succinate dehydrogenase
14	P09382	GAL1 (Galectin-1)
15	O94811	Tubulin polymerization-promoting protein
16	P36873	Serine/threonine-protein phosphatase PP1-gamma catalytic subunit
17	O75925	PIAS1
18	P21980	Protein-glutamine gamma-glutamyltransferase 2
19	P05230	FGF1 (Fibroblast growth factor 1)
20	P05455	La protein (Lupus La protein)
21	Q9P2J5	Leucine–tRNA ligase, cytoplasmic
22	P00734	Thrombin protease (Prothrombin)
23	P15144	CD13 (Aminopeptidase N)
24	Q9UKX7	Nuclear pore complex protein Nup50
25	P00505	Aspartate aminotransferase, mitochondrial
26	O75390	Citrate synthase, mitochondrial
27	Q16637	Survival motor neuron protein
28	P15531	Nucleoside diphosphate kinase A
29	Q12904	Aminoacyl tRNA synthase complex
30	P01033	Erythroid-potentiating activity
31	P06396	Gelsolin
32	P62263	S14
33	P60174	Triose phosphate isomerase
34	P35268	L22 (60S ribosomal protein L22)
35	P62913	60S ribosomal protein L11
36	Q00610	Clathrin
37	P00813	Adenosine Deaminase
38	P22090	S4 (40S ribosomal protein S4, Y isoform 1)
39	P46060	Ran GTPase-activating protein 1
40	P49790	Nuclear pore complex protein Nup153
41	P41182	Bcl6 B-cell lymphoma 6 protein isoform 1
42	O95396	Adenylyltransferase and sulfurtransferase MOCS3
43	Q10589	Bone marrow stromal antigen 2
44	Q9NQX3	Gephyrin, protein anchor
45	P45880	Voltage-dependent anion-selective channel protein 2
46	O95997	Securin
47	P07814	Bifunctional glutamate/proline–tRNA ligase
48	P35222	Catenin beta-1
49	P23396	S3
50	Q5SW96	Low density lipoprotein receptor adapter protein 1
51	P60709	Actin
52	P02511	alpha-crystallin B chain
53	P63104	14–3-3 protein zeta/delta
54	P08183	P-Glycoprotein (transporter)
55	Q13485	Mothers against decapentaplegic homolog 4
56	P02489	alpha-crystallin A chain
57	Q9Y5Y2	Cytosolic Fe-S cluster assembly factor NUBP2
58	P36897	TGF-beta receptor type-1
59	P08253	72 kDa type IV collagenase
60	P50747	Biotin–protein ligase
61	Q9H8X2	Inositol-pentakisphosphate 2-kinase
62	P02730	Band-3 anion exchanger
63	P04637	p53
64	P53384	Cytosolic Fe-S cluster assembly factor NUBP1
65	P49327	Fatty acid synthase
66	P18124	L7
67	Q7Z4W1	L-xylulose reductase
68	O00337	Sodium/nucleoside cotransporter 1
69	Q9H211	DNA replication factor Cdt1
70	O14786	Neuropilin (VEGF receptor)
71	Q6YP21	Kynurenine–oxoglutarate transaminase 3
72	Q86XF0	Bifunctional dihydrofolate reductase-thymidylate synthase
73	P52948	Nuclear pore complex protein Nup98-Nup96
74	O43541	Mothers against decapentaplegic homolog 6
75	O15105	Mothers against decapentaplegic homolog 7
76	P99999	Cytochrome c
77	P46777	Ribosomal protein L5
78	P68104	Elongation factor 1-alpha 1
79	P09429	High mobility group protein B1
80	P04406	Glyceraldehyde-3-phosphate dehydrogenase
81	P07996	Thrombospondin 1
82	P29474	Endothelial nitric oxide synthase
83	P50570	Dynamin
84	P12830	E-cadherin
85	P62277	40S ribosomal protein S13
86	P27635	L10 ribosomal protein
87	Q99217	Amelogenin
88	P19367	Hexokinase
89	P52789	Hexokinase-2
90	Q96CW1	Clathrin adaptor AP-2
91	P15336	ATF2 protein (cyclic AMP-dependent transcription factor)
92	P61254	60S ribosomal protein L26

in the present scenario [8,9]. As organisms evolved towards greater complexity, there was not a proportional increase in the size of their genomes. The fundamental genetic code remained unchanged, meaning that the same genes continued to express proteins. However, the emergence of posttranslational modifications (PTMs) in eukaryotes introduced a form of "multi-tasking" into the equation [10]. Two primary strategies for achieving this multi-functionality were protein promiscuity and moonlighting. Moonlighting proteins highlight versatility of evolution and give rise to many interesting evolutionary questions [11–14]. It is important to note that PTMs themselves played a role in enabling promiscuity and moonlighting, functioning as sub-strategies within these broader concepts [15]. In addition to PTMs, two other notable sub-strategies contributing to this multi-functionality were protein plasticity and intrinsic disorder [15]. Among these sub-strategies, protein promiscuity stood out as a particularly potent mechanism for achieving multi-tasking. This encompassed various forms of promiscuity, such as substrate promiscuity, catalytic promiscuity, and promiscuous protein-protein interactions, all of which played a crucial role in expanding the functional repertoire of proteins [16]. Moonlighting proteins exhibit the remarkable ability to carry out multiple distinct functions autonomously, often without segregating these functions into separate domains within the protein structure [12]. Notably, enzymes serve as striking examples of this phenomenon. In addition to their primary catalytic roles, enzymes are also engaged in entirely unrelated processes, such as autophagy, protein transport, or DNA maintenance [1].

Studies underscore the substantial implications of moonlighting proteins in the context of various diseases and disorders [17–19]. Notably, while the genomes of many organisms likely harbor a wealth of moonlighting proteins with pivotal roles in pathways and disease etiology, the current roster of confirmed moonlighting proteins remains insufficient to provide a comprehensive vista of the intricate cellular mechanisms that underlie their multifaceted functionality [17].

This quantitative shortfall is primarily attributed to the serendipitous nature by which these proteins' additional functions are often discovered, emerging incidentally in unrelated experimental contexts [20]. Therefore, the adoption of a systematic quantitative informatics approach holds the potential to deliver significant contributions by identifying hitherto unknown moonlighting proteins and shedding light on the intricate functional attributes inherent to this intriguing class of proteins [21]. This study presents a quantitative phylogenetic relationship among various human moonlighting proteins based on quantiative

Table 2
List of Moonlighting proteins in humans with their associated Uniprot ID (hyperlinked with respective Uniprot webpage).

Serial No.	Uniprot ID	Moonlighting proteins of human
93	P18074	ERCC2 - TFIIH basal transcription factor complex helicase XPD subunit
94	P27487	CD26/DPPIV
95	Q9BPW9	hRoDH-E2 (dehydrogenase/reductase SDR family member 9)
96	Q9Y697	Cysteine desulfurase, mitochondrial
97	P27797	Calreticulin
98	P40926	Malate dehydrogenase, mitochondrial
99	P54886	Delta-1-pyrroline-5-carboxylate synthase
100	Q16236	Nuclear factor erythroid 2-related factor 2
101	Q15797	Mothers against decapentaplegic homolog 1
102	P07900	HSP90-alpha (heat shock protein HSP 90-alpha)
103	P08238	Heat shock protein HSP 90-beta
104	P09960	Leukotriene A4 hydrolase
105	Q9H009	Nascent polypeptide-associated complex subunit alpha-2
106	P30041	Peroxiredoxin-6
107	Q99988	Growth/differentiation factor 15
108	O95400	CD2BP2 (CD2 antigen cytoplasmic tail-binding protein 2)
109	P14324	Farnesyl pyrophosphate synthase
110	P23381	Tryptophan–tRNA ligase, cytoplasmic
111	P46531	Neurogenic locus notch homolog protein 1
112	P05388	P0
113	Q9Y3E5	Pth2/Bit1 peptidyl-tRNA hydrolase 2/Bcl-2 inhibitor of transcription 1
114	P04818	Thymidylate synthase
115	P04075	Fructose-bisphosphate aldolase A
116	P13716	Delta-aminolevulinic acid dehydratase
117	P13569	CFTR chloride channel (Cystic fibrosis transmembrane conductance regulator)
118	P09622	DLD (dihydrolipoyl dehydrogenase, mitochondrial)
119	P32856	Epimorphin, syntaxin
120	P49759	Dual specificity protein kinase CLK1
121	Q15118	Pyruvate dehydrogenase (acetyl-transferring) kinase isozyme 1, mitochondrial
122	P47992	Chemokine lymphotactin
123	P0C0L4	Complement C4-A
124	P15291	Beta-1,4-galactosyltransferase 1
125	P00533	Epidermal growth factor receptor
126	P49792	E3 SUMO-protein ligase RanBP2
127	P53350	PLK1 (serine/threonine-protein kinase)
128	P05089	Arginase EC:3.5.3.1
129	Q5BJF6	Outer dense fiber protein 2
130	P62328	Thymosin beta-4 (sequester actin)
131	P38936	P21
132	P16402	Histone H1
133	P37198	Nuclear pore glycoprotein p62
134	P49841	Glycogen synthase kinase-3 beta
135	P09104	Gamma-enolase
136	P06733	Alpha-enolase
137	P14618	Pyruvate kinase PKM
138	P62829	60S ribosomal protein L23
139	P36969	Phospholipid hydroperoxide glutathione peroxidase, mitochondrial
140	P00558	Phosphoglycerate kinase
141	P21399	Cytoplasmatic aconitate hydratase/IRP1
142	Q96HA1	Nuclear envelope pore membrane protein POM 121
143	P18754	Regulator of chromosome condensation
144	Q9H2X9	Solute carrier family 12 member 5
145	P23219	Cyclooxygenase-1
146	P84022	Mothers against decapentaplegic homolog 3
147	Q15796	Mothers against decapentaplegic homolog 2
148	P00966	Argininosuccinate synthase
149	O96007	Molybdopterin synthase catalytic subunit
150	Q9Y6I3	Epsin
151	P07355	Annexin-2
152	P53597	Succinyl-coA synthetase (succinate thiokinase)
153	Q15311	RLIP76 (RalA-binding protein 1 - RALBP1)
154	P55084	Trifunctional enzyme subunit beta, mitochondrial
155	P34896	Serine hydroxymethyltransferase, cytosolic
156	P33316	Deoxyuridine 5′-triphosphate nucleotidohydrolase, mitochondrial
157	Q15723	ETS-related transcription factor
158	O60568	Lysyl hydroxylase isoform 3
159	P15559	NAD(*P*)H dehydrogenase [quinone] 1
160	P09038	FGF2 (fibroblast growth factor 2)
161	P19338	Nucleolin
162	P62937	Peptidyl-prolyl cis-trans isomerase A
163	P13010	Ku70/Ku80
164	Q9UQE7	SMC3 (Sister chromatin cohesion)
165	P07954	Fumarate hydratase

sequence based informatics. This investigation further focused on the complex phenomena between sequence based homology and moonlighting functions of proteins. The research raises thought-provoking questions and proposes hypotheses related to the multifaceted functions of moonlighting proteins. These findings present a promising foundation for forthcoming investigations in the realm of moonlighting proteomics.

2. Data acquisition

In the present study, 165 promiscuous proteins/moonlighting proteins of human origin were extracted from the database MultitaskProtDBII (Table 1 & Table 2). The list of all these protein sequences including their Uniport ID is supplied as Supplementary file-1.

3. Methods

3.1. Composition profiler

The Composition Profiler was used to generate an amino acid composition profile of all the human moonlighting proteins analyzed in this study [22]. This set of amino acid sequences was the query set and the 'Protein Data Bank Select 25' was the background set. We also generated a composition profile for experimentally validated disordered proteins from the DisProt database. The generated profiles represent plots showing normalized enrichment or depletion of a given residue calculated as $\frac{(C_x - C_{order})}{C_{order}}$, where C_x is the content of a given residue in its query protein, and C_{order} is the content of the same residue in the PDB Select 25.

3.2. Determining amino acid frequency composition

Count of every amino acid in a sequence, termed amino acid frequency, was computed for all 165 moonlighting protein [23–25]. Due to varied lengths of moonlighting protein sequences, percentage of amino acids in a sequence (obtained from dividing the amino acid frequencies by the length of that sequence and multiplied by 100) is termed the relative frequency of amino acids in that sequence. Relative frequency of 20 amino acids represents a 20 dimensional vector for each protein sequence. Further, 165 moonlighting protein sequences resulted into a 165 dimensional vector for each amino acid. Correlation coefficient matrix (20×20) based on all possible pairs of amino acids was computed from the relative frequency vectors of amino acids.

3.2.1. Evaluating Shannon entropy of sequences

Shannon entropy (SE) is a measure of the information content in a system [26]. SE of each moonlighting protein sequence is evaluated by the formula:

$$SE = -\sum_{i=1}^{20}(p_i^{*} log_2 p_i)$$

p_i is the count of amino acid i in that sequence divided by the length of that sequence [23]. *SE* reflects the degree of randomness in the amino

Table 3
List of features of dimension 1182 calculated by I-features.

List of various features (1182 dimensional) calculated by iFeature		
Descriptor groups	Descriptor	Dimension
Amino acid composition	Amino acid composition (AAC)	20
	Dipeptide composition (DPC)	400
Grouped amino acid composition	Grouped amino acid composition (GAAC)	5
	Grouped dipeptide composition (GDC)	25
Autocorrelation	Moran autocorrelation	240
C/T/D	C/T/D composition (CTDC)	39
Conjoint triad	Conjoint triad	343
Quasi-sequence-order	Sequence-order-coupling number (SoC number)	60
Pseudo-amino acid composition	Pseudo-amino acid composition (PAAC)	50

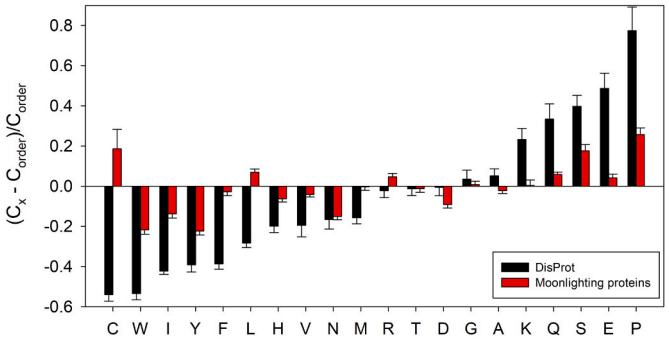


Fig. 1. Amino acid composition profile of 165 human moonlighting proteins (red bars). The fractional difference is calculated as $\frac{(C_x - C_{order})}{C_{order}}$, where C_x is the content of a given amino acid in the query set (165 human moonlighting proteins or known intrinsically disordered proteins), and C_{order} is the content of a given amino acid in the background set (Protein Data Bank Select 25). The amino acid residues are ranked from most order-promoting residue to most disorder-promoting residue. Positive values indicate enrichment, and negative values indicate depletion of a particular amino acid. The composition profile generated for experimentally validated disordered proteins from the DisProt database (black bars) is shown for comparison. In both cases, error bars correspond to standard deviations over 10,000 bootstrap iterations. The composition profile analysis showed that 6 out of 10 order-promoting residues (W, I, Y, H, V, and N) were statistically depleted in the moonlighting proteins (p-value ≤ 0.05), and 5 out of 10 disorder-promoting residue (R, Q, S, E, and P) were significantly enriched in these proteins. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

acids count in a given sequence. A higher value of *SE* indicates greater diversity, while a lower value indicates less diversity.

3.3. Determining homogeneous poly-string frequency of amino acids

Homogeneous poly-string of length n is defined as n consecutive occurrence of a particular amino acid [27]. For example, *AAATTAATT* represents one homogeneous poly-string of A with length 3, one homogeneous poly-string of A with length 2, and two homogeneous poly-strings of T with length 2. Note that while counting the number of a homogeneous poly-string of length n, only the exclusive/exact occurrence of length n would be taken into consideration. The maximum of lengths of homogeneous poly-strings considering all amino acids across all sequences was computed and accordingly, counts of homogeneous poly-strings of all possible lengths (starting from 1 to maximum length) for each amino acid present in a given protein sequence was enumerated.

3.4. Evaluating polar, non-polar residue profiles

Every amino acid in a given moonlighting protein sequence was identified as polar(P) or non-polar(Q). Thus, every protein sequence became a binary sequence with two symbols: P and Q. Through this binary P-Q profile, a spatial assembly of polar and non-polar residues over a protein sequence was depicted [28–30].

3.4.1. Change response sequences based on polar, non-polar residue profiles

There are four possible changes between two consecutive residues of a polar and non-polar profile, namely Polar to Polar (PP), Polar to Non-polar (PN), Non-polar to Non-Polar (NN) and Non-polar to Polar (NP). Such changes were accounted in form of a sequence according to the spatial sequential arrangement of polar, non-polar residues in a given P-Q binary profile. We call this sequence "P-Q Change Response Sequence (CRS_{PQ})". Frequency of each of the four changes was enumerated from binary polar non-polar profile corresponding to each moonlighting protein.

3.5. Evaluating acidic, basic, neutral residue profiles

Every amino acid in a given moonlighting protein sequence was identified as acidic (A), basic (B), and neutral (N). Thus, every protein sequence became a ternary valued sequence (A-B-N profiles) with three symbols: A, B, and N.

3.5.1. Change response sequences based on acidic-basic-neutral residue profiles

There are nine possible changes between two consecutive residues of an A-B-N profile, namely Acidic to Acidic (AA), Acidic to Basic (AB), Acidic to Neutral (AN), Basic to Acidic (BA), Basic to Basic (BB), Basic to Neutral (BN), Neutral to Acidic (NA), Neutral to Basic (NB), Neutral to Neutral (NN). Such changes were accounted in form of a sequence according to the spatial sequential arrangement of acidic, basic and neutral residues in a given A-B-N ternary profile. We designate this sequence "A-B-N Change Response Sequence (CRS_{ABN})". Frequency of each of the nine changes was counted for a given ternary A-B-N profile corresponding to each moonlighting protein.

3.6. Evaluating intrinsic protein disorder

Predisposition for intrinsic disorder of all the human moonlighting proteins analyzed in this study were determined using a set of commonly used per-residue disorder predictors, such as PONDR® VLS2, PONDR® VL3, PONDR® VLXT, PONDR® FIT, IUPred-Long, IUPred-Short [31–36]. A web platform called Rapid Intrinsic Disorder Analysis Online (RIDAO) was used to gather results from each predictor in bulk [37]. The percent of predicted intrinsically disorder residues (PPIDR) for each protein was used to classify each protein based on their level of disorder. A residue was considered to be disordered if it had a value of 0.5 or higher. Generally, a PPIDR value of less than 10 % is taken to correspond to a highly ordered protein, PPIDR between 10 % and 30 % is ascribed to moderately disordered protein, and PPIDR greater than 30 % corresponds to a highly disordered protein [38,39]. In addition to PPIDR, mean disorder score (MDS) was calculated for each query protein as a protein length-normalized sum of all the per-residue disorder scores. The per-residue disorder score ranges from 0 to 1, where a score of 0 indicates fully ordered residues and a score of 1 indicates fully disordered residues. Residues with scores above the threshold of 0.5 were considered *disordered residues*. Residues with disorder scores between 0.25 and 0.5 were categorized as *highly flexible*, while those with scores between 0.1 and 0.25 were classified as *moderately flexible* [36]. We also utilized two binary predictors of disorder, the charge-hydropathy (CH) plot and the cumulative distribution function (CDF) (both accessible at Predictor of Natural Disordered Regions), to assess intrinsic disorder at the whole protein level [40,41]. These binary

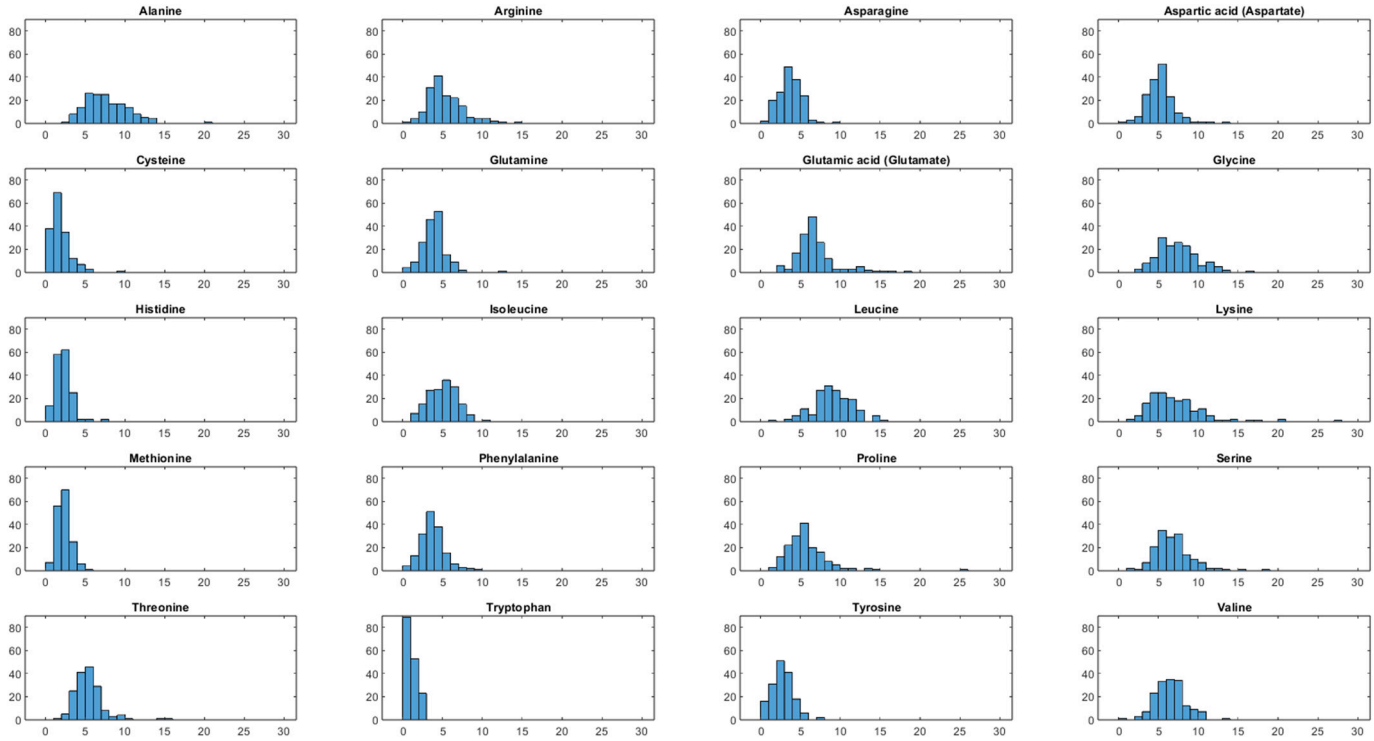


Fig. 2. Histogram of relative frequency of each amino acid. X axis denotes percentage and Y axis denotes number of sequences.

predictors can be combined to generate CH-CDF plots that distinguish hybrid proteins containing moderate order and disorder from entirely ordered or entirely disordered proteins [42,43]. Herein, we utilized RIDAO generated CH-CDF plots based on VLXT CDF scores.

3.6.1. Change response sequences based on intrinsic protein disorder residues

There are sixteen possible changes between two consecutive residues where residues were denoted as disordered (D), highly flexible (HF), moderately flexible (MF) and other (O) in moonlighting protein sequences namely disordered to disordered (D_D), disordered to highly flexible (D_HF), disordered to moderately flexible (D_MF), disordered to other (D_O) and similarly rest twelve (HF_D, HF_HF HF_MF, HF_O, MF_D, MF_HF, MF_MF, MF_O, O_D, O_HF, O_MF, and O_O). Such changes were accounted in the form of a sequence according to the spatial sequential arrangement of D, HF, MF, and O residues for each moonlighting protein sequence. Frequency of each of the sixteen changes was counted from these change response sequences.

3.7. Structural and physicochemical features (I-features)

Structural and physicochemical descriptors extracted from sequence data have been widely used to characterize sequences and predict structural, functional, expression and interaction profiles of proteins. iFeature (I-features), a versatile Python-based toolkit was deployed to extract structural and physicochemical properties of the moonlighting proteins [44].

A list of 1182 features which were deployed in the present study was summarized in Table 3 [44].

3.8. Formation of distance matrices and dendrograms

Euclidean distance was evaluated between feature vectors of all pairs of moonlighting protein sequences, for each of the following five features: relative frequency of amino acids (dimension 20), relative frequency of changes obtained from polar-nonpolar profiles (dimension 4), relative frequency of changes obtained from acidic-basic-neutral profiles (dimension 9), relative frequency of changes obtained from disordered, highly flexible, moderately flexible, and other residues (dimension 16), structural and physicochemical features (dimension 1182) [45].

Table 4
Moonlighting sequences with significantly high percentage (≥15%) of certain amino acids.

Uniprot ID	Alanine	Glutamic acid	Glycine	Leucine	Lysine	Proline	Serine	Threonine
P49790							18.1	
O95396				15				
P99999					17.14			
P09429		16.74			20			
Q99217						25.65		
P61254					16.55			
P62328		18.18			20.45			
P16402	20.81				27.6			
P37198								15.52
Q96HA1							15.21	
Q15311		15.57						
P09038			16.32					

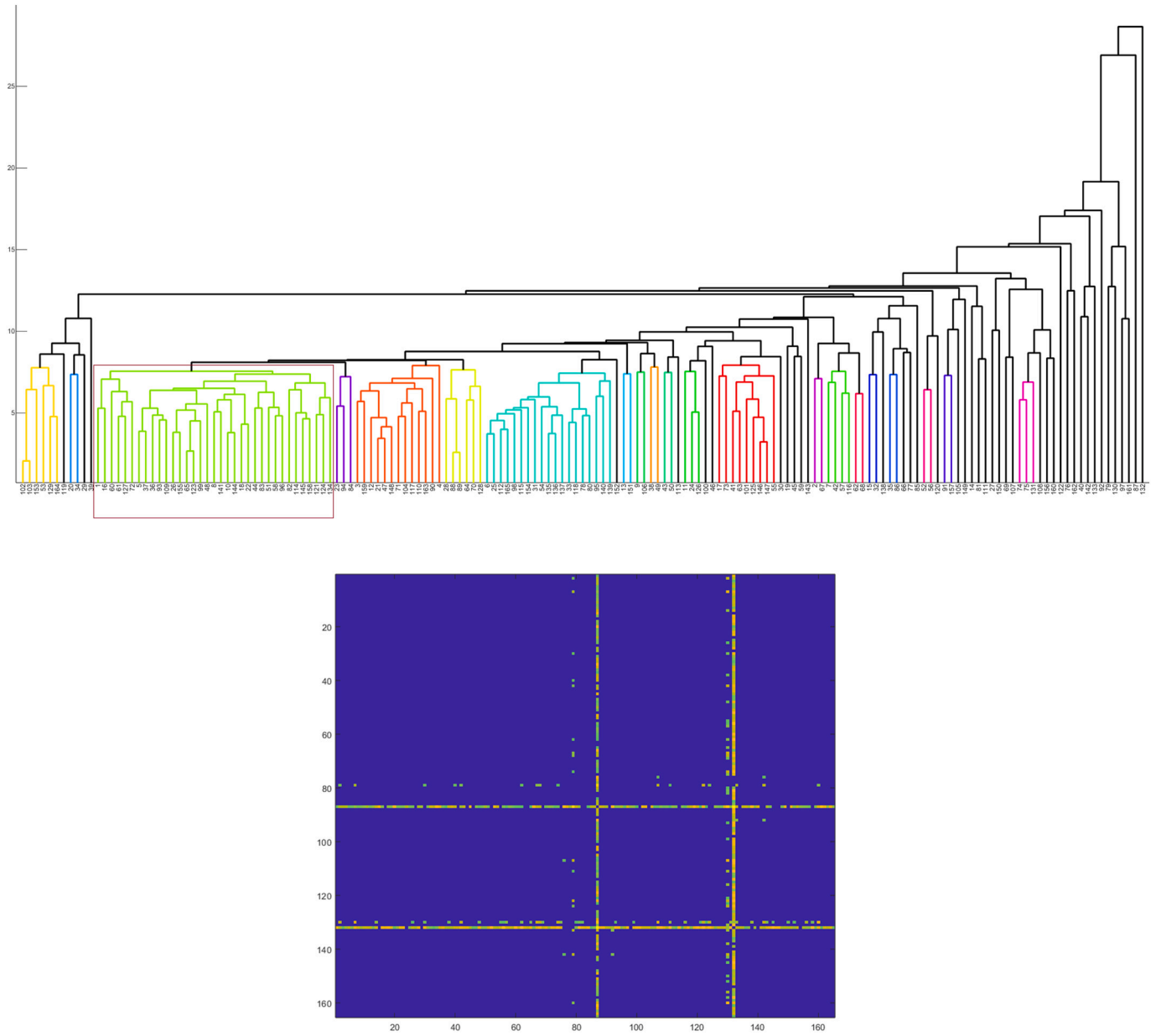


Fig. 3. (A): Phylogenetic relationship among the moonlighting proteins based on relative frequency of amino acids. (B): Distance matrix consisting distances between pair of amino acid frequency vectors corresponding to each moonlighting protein (distance less than 25 considered to be zero).

Relative frequency was defined as the frequency divided by the length of the sequence and multiplied by 100. Each of the 1182 structural and physicochemical features was normalized in the range of 0 to 100. Each of the five features produces a distance matrix of dimension 165×165. For better visualization, distances less than a threshold (empirically chosen separately for each of the distance matrices) were converted to zero. Dendrograms were formed based on Euclidean distances and average linkage. Different color thresholds (empirically chosen) were used in different dendrograms. If the color threshold had the value T, then each group of nodes whose linkage was less than T was assigned a unique color in the dendrogram.

3.9. Cumulative clustering of moonlighting proteins

Each of the five distance matrices (as mentioned in Section 3.7) was normalized by dividing each entry of the matrix by the maximum of that matrix and multiplied by 100. Then all five normalized matrices were added to form a distance matrix for cumulative clustering.

3.10. Agglomerated distal-proximal relationships of moonlighting proteins

A moonlighting protein sequence is termed as '*outlier*' with respect to a feature, only if the feature vector of the sequence is distant from most of the feature vectors of the remaining sequences. Set of outlier moonlighting proteins are termed here as '*distal*' proteins.

A set of moonlighting protein sequences was called as '*proximal set*' with respect to a feature if the sequences belong to the same cluster or immediate adjacent cluster in the dendrogram formed from that feature. Clusters formed from the cumulative clustering were analyzed in detail with respect to each feature to infer how the features individually contributed to the proximal relationship among sequences.

It is worth mentioning that feature extractions and subsequent

Table 5
Clusters derived from relative frequency of amino acids.

Clusters	Sequences
Cluster-1	{102,103,153,53,129,164}
Cluster-2	{20,34}
Cluster-3	{1,16,60,61,127,72,5,37,36,93,109,26,155,65,123,99,48,8141,10,144,18,22,44,83,51,58,96,82,114,145,158,121,124,134}
Cluster-4	{23,94,84}
Cluster-5	{3159,12,21,47,148,71,104,117,110,163,90,4}
Cluster-6	{28,88,89,64,70,128}
Cluster-7	{6,25,112,165,98,115,154,31,54,135,136,137,33,118,78,80,95,140,139}
Cluster-8	{13,151}
Cluster-9	{9106}
Cluster-10	{38,49}
Cluster-11	{43,50}
Cluster-12	{11,24,126}
Cluster-13	{17,73,41,63,101,125,146,147,55}
Cluster-14	{2,67}
Cluster-15	{7,42,57,116}
Cluster-16	{62,68}
Cluster-17	{15,32}
Cluster-18	{35,86}
Cluster-19	{52,56}
Cluster-20	{91,157}
Cluster-21	{74,75,131}

computations were carried out in *MATLAB*(2023b).

4. Results and analyses

4.1. Similitude and dissimilitude of moonlighting proteins based on amino acid frequency

4.1.1. Compositional profile of human moonlighting proteins

It is known that the amino acid compositions of ordered and intrinsically disordered proteins are characterized by noticeable difference. In fact, the disordered proteins/regions were shown to be significantly depleted in bulky hydrophobic (I, L, and V) and aromatic amino acid residues (W, Y, F, and H) that are often involved in the formation of the hydrophobic core of a folded globular protein. Disordered proteins/regions also show a low content of C, N, and M residues. These residues, C, W, I, Y, F, L, H, V, N, and M, which are depleted in disordered proteins and regions are defined as order-promoting amino acids. On the other hand, disordered proteins and regions are substantially enriched in disorder-promoting amino acids, such as R, T, D, G, A, K, Q, S, E, and P [22,35,46–48]. These biases in the amino acid composition can be

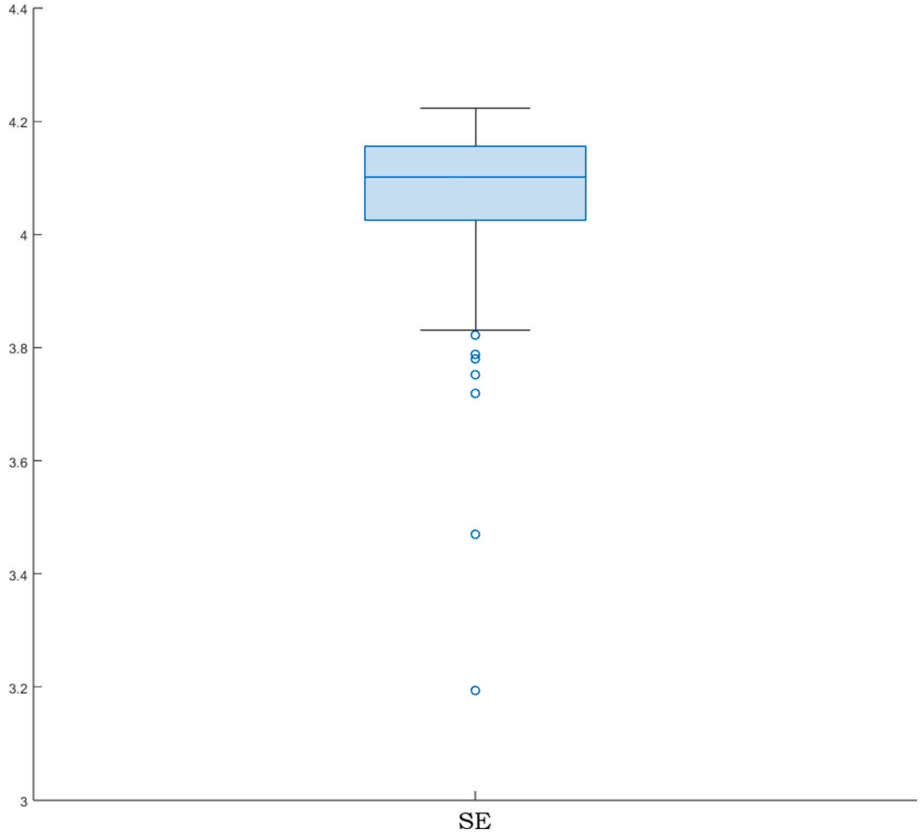


Fig. 4. Boxplot of Shannon entropy based on amino acid frequencies derived from respective moonlighting proteins.

Table 6
Frequency of homogeneous poly-string of length 1, 2, 14 doe each moonlighting proteins.

Uniprot ID	Homogenous polystring of length l										Uniprot ID	Homogenous polystring of length l									
	l = 1	l = 2	l = 3	l = 4	l = 5	l = 6	l = 7	l = 9	l = 10	l = 14		l = 1	l = 2	l = 3	l = 4	l = 5	l = 6	l = 7	l = 9	l = 10	l = 14
P28482	327	12	1	0	0	1	0	0	0	0	P12830	746	51	8	0	2	0	0	0	0	0
Q99714	232	13	1	0	0	0	0	0	0	0	P62277	137	7	0	0	0	0	0	0	0	0
P55263	306	22	4	0	0	0	0	0	0	0	P27635	202	6	0	0	0	0	0	0	0	0
Q7L0Y3	340	24	5	0	0	0	0	0	0	0	Q99217	143	21	2	0	0	0	0	0	0	0
P47897	681	44	2	0	0	0	0	0	0	0	P19367	796	59	1	0	0	0	0	0	0	0
Q02252	469	28	2	1	0	0	0	0	0	0	P52789	813	52	0	0	0	0	0	0	0	0
P19971	404	36	2	0	0	0	0	0	0	0	Q96CW1	384	24	1	0	0	0	0	0	0	0
P06744	504	27	0	0	0	0	0	0	0	0	P15336	414	39	3	1	0	0	0	0	0	0
P62826	187	13	1	0	0	0	0	0	0	0	P61254	123	11	0	0	0	0	0	0	0	0
P32189	487	33	2	0	0	0	0	0	0	0	P18074	686	31	4	0	0	0	0	0	0	0
P78545	320	19	3	1	0	0	0	0	0	0	P27487	675	41	3	0	0	0	0	0	0	0
Q15046	509	38	4	0	0	0	0	0	0	0	Q9BPW9	283	15	2	0	0	0	0	0	0	0
P21912	237	17	3	0	0	0	0	0	0	0	Q9Y697	389	28	4	0	0	0	0	0	0	0
P09382	129	3	0	0	0	0	0	0	0	0	P27797	356	22	3	2	0	0	0	0	0	0
O94811	190	13	1	0	0	0	0	0	0	0	P40926	293	18	3	0	0	0	0	0	0	0
P36873	273	22	2	0	0	0	0	0	0	0	P54886	679	55	2	0	0	0	0	0	0	0
O75925	546	41	4	0	1	1	0	0	0	0	Q16236	534	31	3	0	0	0	0	0	0	0
P21980	601	40	2	0	0	0	0	0	0	0	Q15797	395	28	2	2	0	0	0	0	0	0
P05230	133	11	0	0	0	0	0	0	0	0	P07900	606	47	6	2	0	1	0	0	0	0
P05455	352	28	0	0	0	0	0	0	0	0	P08238	599	48	7	2	0	0	0	0	0	0
Q9P2J5	1044	55	5	0	0	0	1	0	0	0	P09960	565	23	0	0	0	0	0	0	0	0
P00734	549	35	1	0	0	0	0	0	0	0	Q9H009	190	6	3	1	0	0	0	0	0	0
P15144	843	52	5	0	1	0	0	0	0	0	P30041	190	14	2	0	0	0	0	0	0	0
Q9UKX7	377	38	5	0	0	0	0	0	0	0	Q99988	271	14	3	0	0	0	0	0	0	0
P00505	383	17	3	1	0	0	0	0	0	0	O95400	282	17	7	1	0	0	0	0	0	0
O75390	414	23	2	0	0	0	0	0	0	0	P14324	363	25	2	0	0	0	0	0	0	0
Q16637	230	19	2	0	2	0	0	0	1	0	P23381	417	24	2	0	0	0	0	0	0	0
P15531	138	7	0	0	0	0	0	0	0	0	P46531	2197	145	10	8	0	1	0	0	0	0
Q12904	265	22	1	0	0	0	0	0	0	0	P05388	265	19	3	0	1	0	0	0	0	0
P01033	185	9	0	1	0	0	0	0	0	0	Q9Y3E5	164	6	1	0	0	0	0	0	0	0
P06396	675	46	5	0	0	0	0	0	0	0	P04818	277	18	0	0	0	0	0	0	0	0
P62263	129	11	0	0	0	0	0	0	0	0	P04075	304	24	4	0	0	0	0	0	0	0
P60174	219	15	0	0	0	0	0	0	0	0	P13716	282	21	2	0	0	0	0	0	0	0
P35268	98	9	1	1	1	0	0	0	0	0	P13569	1279	85	9	1	0	0	0	0	0	0
P62913	158	10	0	0	0	0	0	0	0	0	P09622	451	29	0	0	0	0	0	0	0	0
Q00610	1463	97	6	0	0	0	0	0	0	0	P32856	236	21	2	1	0	0	0	0	0	0
P00813	314	20	3	0	0	0	0	0	0	0	P49759	422	29	0	1	0	0	0	0	0	0
P22090	245	9	0	0	0	0	0	0	0	0	Q15118	393	20	1	0	0	0	0	0	0	0
P46060	490	25	5	0	0	3	0	0	0	1	P47992	102	6	0	0	0	0	0	0	0	0
P49790	1181	98	17	5	3	2	0	0	0	0	P0C0L4	1520	91	10	3	0	0	0	0	0	0
P41182	642	29	2	0	0	0	0	0	0	0	P15291	337	24	3	1	0	0	0	0	0	0
O95396	386	28	6	0	0	0	0	0	0	0	P00533	1078	62	1	0	1	0	0	0	0	0
Q10589	139	16	3	0	0	0	0	0	0	0	P49792	2801	184	14	2	1	0	0	0	0	0
Q9NQX3	630	43	2	2	0	1	0	0	0	0	P53350	517	37	4	0	0	0	0	0	0	0
P45880	270	12	0	0	0	0	0	0	0	0	P05089	287	16	1	0	0	0	0	0	0	0
O95997	174	14	0	0	0	0	0	0	0	0	Q5BJF6	724	42	4	1	1	0	0	0	0	0
P07814	1325	84	5	1	0	0	0	0	0	0	P62328	42	1	0	0	0	0	0	0	0	0
P35222	661	48	8	0	0	0	0	0	0	0	P38936	145	8	1	0	0	0	0	0	0	0
P23396	213	12	2	0	0	0	0	0	0	0	P16402	160	26	3	0	0	0	0	0	0	0
Q5SW96	273	14	1	1	0	0	0	0	0	0	P37198	432	34	6	1	0	0	0	0	0	0
P60709	333	16	2	1	0	0	0	0	0	0	P49841	367	22	3	0	0	0	0	0	0	0
P02511	149	13	0	0	0	0	0	0	0	0	P09104	402	13	2	0	0	0	0	0	0	0
P63104	201	22	0	0	0	0	0	0	0	0	P06733	396	16	2	0	0	0	0	0	0	0
P08183	1095	83	5	1	0	0	0	0	0	0	P14618	483	24	0	0	0	0	0	0	0	0
Q13485	482	25	5	0	1	0	0	0	0	0	P62829	122	9	0	0	0	0	0	0	0	0
P02489	155	9	0	0	0	0	0	0	0	0	P36969	178	9	0	0	0	0	0	0	0	0
Q9Y5Y2	237	14	2	0	0	0	0	0	0	0	P00558	361	22	4	0	0	0	0	0	0	0
P36897	429	29	1	1	0	0	0	1	0	0	P21399	779	52	2	0	0	0	0	0	0	0
P08253	583	29	5	1	0	0	0	0	0	0	Q96HA1	932	109	22	7	1	0	0	0	0	0
P50747	638	38	4	0	0	0	0	0	0	0	P18754	373	24	0	0	0	0	0	0	0	0
Q9H8X2	442	23	1	0	0	0	0	0	0	0	Q9H2X9	974	72	7	0	0	0	0	0	0	0
P02730	748	68	9	0	0	0	0	0	0	0	P23219	524	31	3	1	0	0	0	0	0	0
P04637	311	32	6	0	0	0	0	0	0	0	P84022	370	26	1	0	0	0	0	0	0	0
P53384	274	23	0	0	0	0	0	0	0	0	Q15796	401	30	2	0	0	0	0	0	0	0
P49327	2163	156	12	0	0	0	0	0	0	0	P00966	386	13	0	0	0	0	0	0	0	0
P18124	213	13	3	0	0	0	0	0	0	0	O96007	160	11	2	0	0	0	0	0	0	0
Q7Z4W1	220	12	0	0	0	0	0	0	0	0	Q9Y6I3	453	51	7	0	0	0	0	0	0	0
O00337	553	43	2	1	0	0	0	0	0	0	P07355	314	11	1	0	0	0	0	0	0	0
Q9H211	465	34	3	1	0	0	0	0	0	0	P53597	279	25	3	2	0	0	0	0	0	0
O14786	832	35	7	0	0	0	0	0	0	0	Q15311	527	55	4	0	0	1	0	0	0	0
Q6YP21	410	19	2	0	0	0	0	0	0	0	P55084	411	25	3	1	0	0	0	0	0	0
Q86XF0	163	9	2	0	0	0	0	0	0	0	P34896	436	20	1	1	0	0	0	0	0	0

(*continued on next page*)

Table 6 (*continued*)

Uniprot ID	Homogenous polystring of length l										Uniprot ID	Homogenous polystring of length l									
	l = 1	l = 2	l = 3	l = 4	l = 5	l = 6	l = 7	l = 9	l = 10	l = 14		l = 1	l = 2	l = 3	l = 4	l = 5	l = 6	l = 7	l = 9	l = 10	l = 14
P52948	1573	100	13	0	1	0	0	0	0	0	P33316	218	17	0	0	0	0	0	0	0	0
O43541	409	28	3	4	0	1	0	0	0	0	Q15723	501	36	4	2	0	0	0	0	0	0
O15105	344	23	7	2	0	0	1	0	0	0	O60568	643	35	4	2	1	0	0	0	0	0
P99999	88	7	1	0	0	0	0	0	0	0	P15559	236	15	1	0	1	0	0	0	0	0
P46777	245	18	4	1	0	0	0	0	0	0	P09038	232	22	4	0	0	0	0	0	0	0
P68104	405	27	1	0	0	0	0	0	0	0	P19338	515	66	10	2	3	0	0	0	1	0
P09429	147	21	3	3	1	0	0	0	0	0	P62937	151	7	0	0	0	0	0	0	0	0
P04406	299	18	0	0	0	0	0	0	0	0	P13010	614	53	4	0	0	0	0	0	0	0
P07996	1027	60	6	0	1	0	0	0	0	0	Q9UQE7	1073	66	4	0	0	0	0	0	0	0
P29474	1046	71	5	0	0	0	0	0	0	0	P07954	445	28	3	0	0	0	0	0	0	0
P50570	736	61	4	0	0	0	0	0	0	0											

visualized using a web-based tool Composition Profiler for semi-automatic discovery of enrichment or depletion of amino acids in query proteins [22]. Results of the corresponding analysis of 165 human moonlighting proteins are shown in Fig. 1. This analysis revealed that out of the ten order-promoting residues six (W, I, Y, H, V, and N) were significantly depleted in the moonlighting proteins, whereas five disorder-promoting residues (R, Q, S, E, and P) were significantly enriched. Curiously, relative to the ordered proteins from the PDB Select 25 set, moonlighting proteins analyzed in this study were significantly enriched in C and L residues while being depleted in D residues.

4.1.2. *Relative frequency of amino acids and their correlation*

It was found that the maximum frequency of leucine was in 63 sequences followed by alanine (23), glycine (19), lysine (18), serine (13), glutamic acid (12), proline (6), valine (4), aspartic acid (3), threonine (3), and arginine (1) (Numeral inside parentheses denotes the number of moonlighting sequences, where the respective amino acid frequency was found to be maximum). The amino acids that had the minimal frequency along the sequences in the decreasing order were tryptophan (94) followed by cysteine (42), histidine (9), methionine (9), tyrosine (4), asparagine (2), arginine (1), and phenylalanine (1).

From the relative frequency vector (consisting of 165 elements) of each amino acid, maximum and minimum standard deviation was found for lysine (3.56) and Tryptophan (0.68) respectively (Fig. 2). Table 4 shows that only eight amino acids (alanine, glutamic acid, glycine, leucine, lysine, proline, serine, and threonine) were present with more than 15 % in twelve sequences.

As noticed that the frequency of tryptophan is least in most of the moonlighting protein sequences. One of the reasons for its scarcity being its indole side chain possessing an aromatic bi-nuclear ring structure, unlike the single-ring aromatics found in phenylalanine, tyrosine, and histidine- allowing extensive hydrophobic pockets to be located that facilitate interactions requiring a relatively substantial surface area. Because of these bulky structural characteristics, the biosynthetic pathway of tryptophan stands out as the most energy-demanding among all amino acids, requiring seventy ATP molecules for its production, in contrast to serine, which only requires ten ATP molecules. Tryptophan is coded by only one codon- UGG. Out of sixty-four codons, only one codon is available for its expression. This absence of ambiguity is another reason behind its infrequent appearance [49].

Leucine is a hydrophobic non-polar aliphatic amino acid with a branched side chain. Leucine emerges as a potential amino acid with the greater inclination for α-helical structure formation. Among Glu, Ala, Leu, and His, which exhibit the highest occurrences within helical regions, Leu stands out as the predominant residue within the central helical cores of proteins. Therefore a high frequency of Leu suggests that

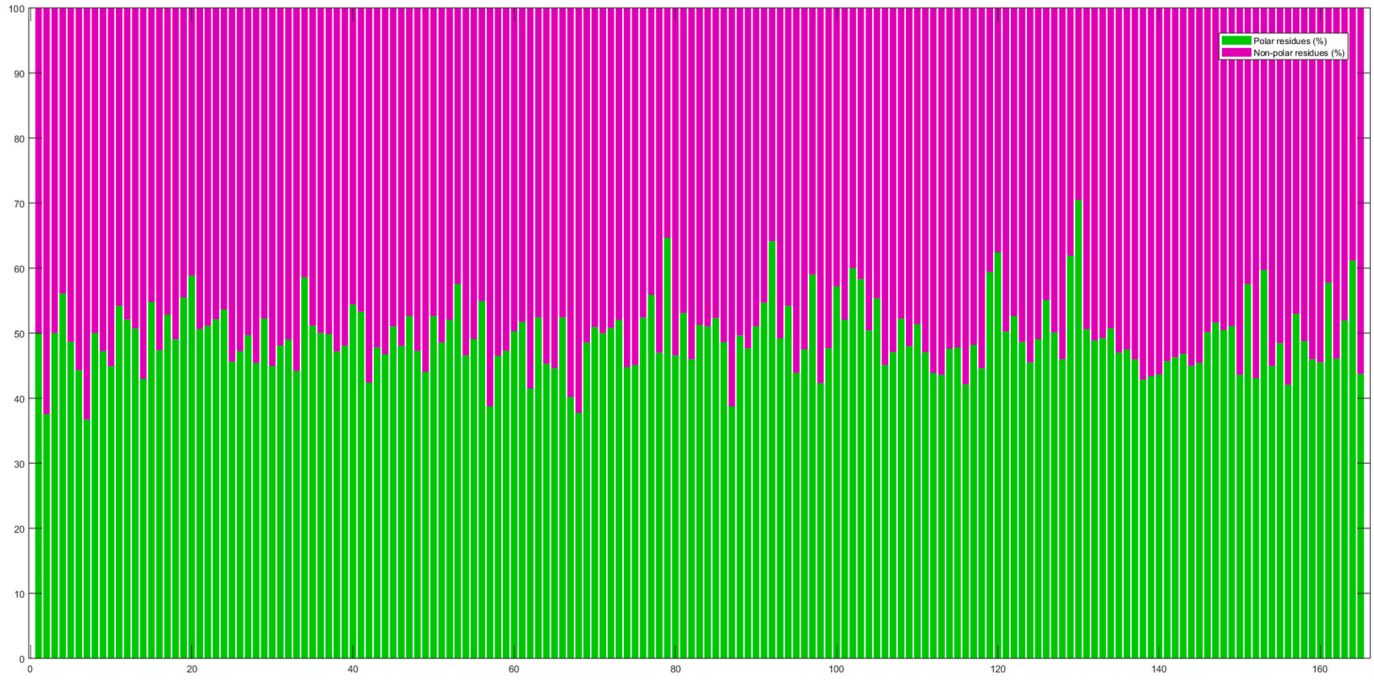


Fig. 5. Percentages of polar, non-polar residues in moonlighting proteins.

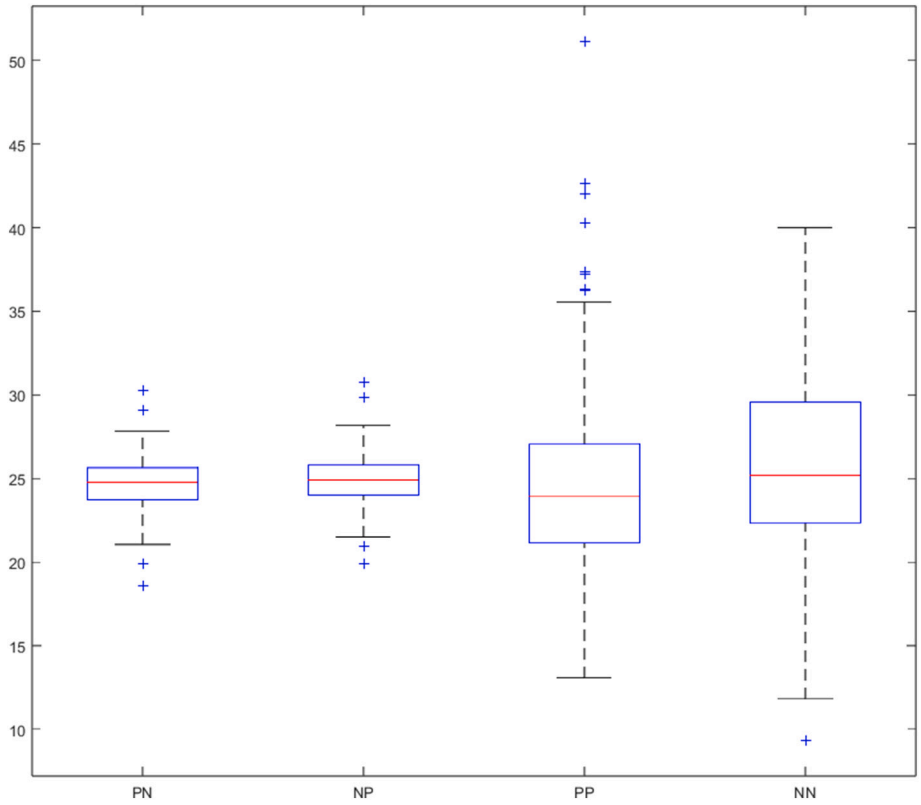


Fig. 6. Box-plot of the relative frequency of PN, NP, PP, and NN changes in moonlighting proteins.

these proteins may have a stable core structure [50]. From the above observations, we can conclude that such a high frequency of amino acids being at the maximal and minimal ends is suggestive of a probable enormous functional diversity. The amino acids present in a higher concentration majorly play a role in key functional regions such as active sites, binding sites, or regions responsible for protein-protein interactions. Also, these amino acids might be preferred due to their ability to form specific secondary structures (e.g., helices or sheets) or specific motifs. In contrast, the amino acids with low frequency conveyed that they are less common and there might be a subset of conserved amino acids that are crucial for the function or structure of the moonlighting protein family.

It was found that absolute values of the Spearman's rank correlation coefficient for all possible pairs of amino acids were less than 0.5. Hence, no significant positive/negative correlation was obtained in frequency composition of amino acids.

4.1.3. *Distance matrix and phylogenetic relationship*

Based on a threshold of distance of 8, 126 moonlighting sequences formed 21 clusters as obtained in the dendrogram (Fig. 3(A)). Largest clusters comprised of 35 sequences as marked by brown rectangular box (Fig. 3(A)) (Table 5). Four sequences namely P16402, Q99217, P62328, and P09429 were more than 25 distance away from 153, 136, 43, and 17 sequences, respectively. Among all possible pairwise distances, maximum distance (38.86) was found between P16402 and Q99217. These findings were displayed in form of a distance matrix with a cut off of 25 (Fig. 3(B)).

4.1.4. *Shannon entropy of amino acid frequency*

Amino acid frequency based Shannon entropy (SE) was calculated for each moonlighting protein sequence. The highest possible SE (4.22) was found for the sequence P84022, whereas the lowest SE (3.19) was obtained for *P*16402. Sequence P62328 also had a much lower SE (3.47) as obtained in Fig. 4. It was noticed that all four sequences namely P16402, P62328, Q99217, and P09429 which were highlighted in Fig. 3 (B) had SEs ranging from 3.19 to 3.75.

The variation in entropy implies the functional versatility and adaptability of moonlighting proteins [26,51]. Proteins with lower entropy (closer to 3.19) may possess a more specific or constrained set of functions, as their amino acid compositions show less variation [51]. On the other hand, proteins with higher entropy (closer to 4.21) may have more flexible or adaptable functional roles, as their amino acid compositions exhibit greater diversity [52]. Furthermore, it was inferred that the wide range of SEs suggests that moonlighting proteins have undergone evolutionary adaptations to acquire their multiple functions. The different entropy levels likely result from various selective pressures acting on these proteins throughout their evolutionary history [53]. Proteins with higher entropy may have experienced more relaxed selection, allowing for greater divergence in their amino acid compositions and the acquisition of novel functions [54]. Despite the wide range, there may be functional constraints that limit the amino acid composition variability within certain moonlighting proteins [55]. Certain regions or residues may be conserved across the set to maintain essential functional properties required for their moonlighting roles [56]. These conserved features would contribute to the lower end of the entropy range [57].

4.2. *Homogeneous poly-string frequency of amino acids in moonlighting proteins*

The analysis unveiled the maximum length of a homogeneous poly-string to be 14, encompassing all amino acids across the moonlighting protein sequences. This prompted the enumeration of the frequency of homogeneous poly-strings with lengths ranging from 1 to 14 for each of the twenty amino acids, as detailed in Supplementary file-2. The cumulative frequency of all homogeneous poly-strings was then calculated for each moonlighting protein and summarized in Table 6. Interestingly, no poly-strings of lengths 8, 11, 12, and 13 were present in any of the

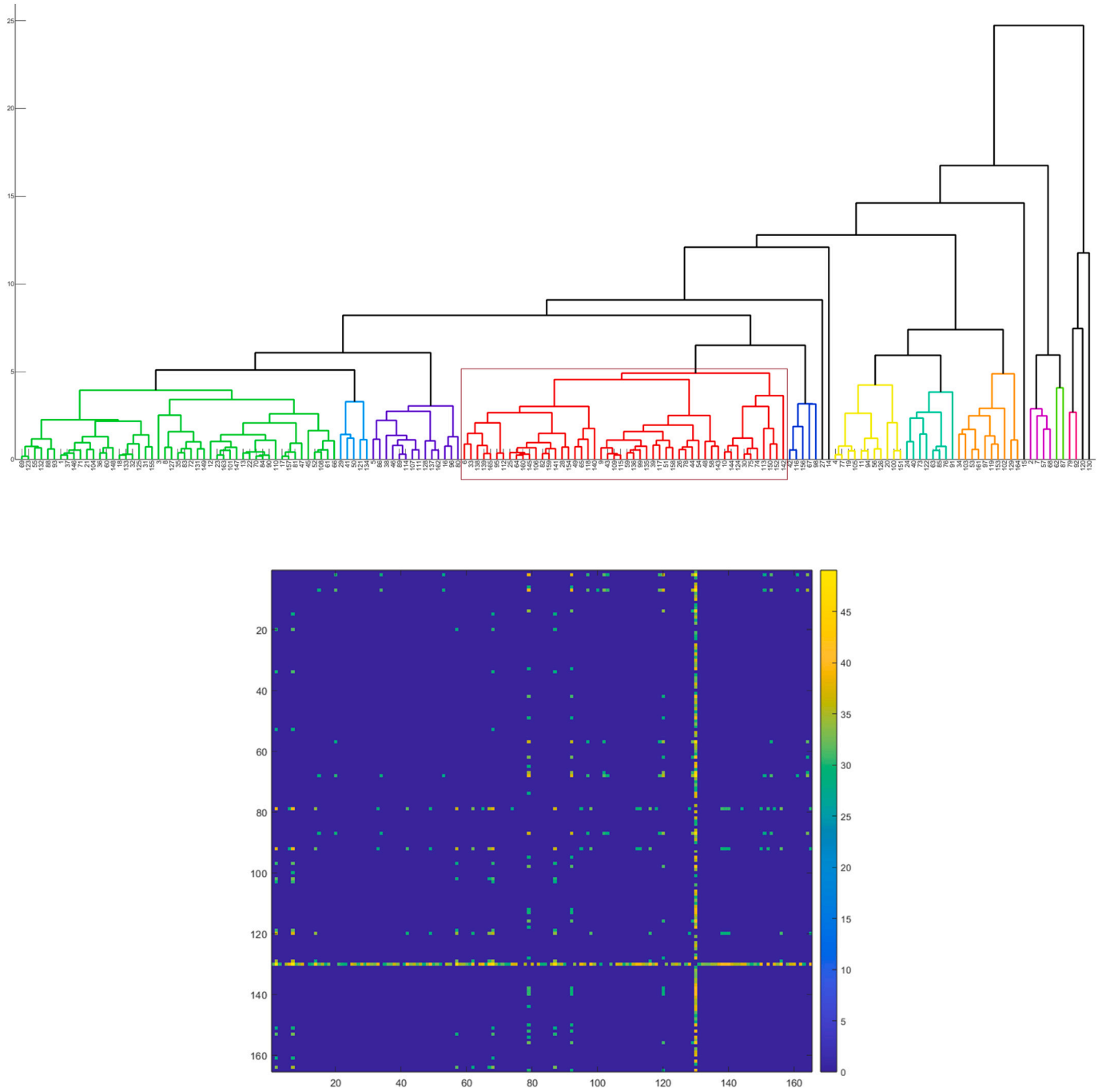


Fig. 7. (A): Phylogenetic relationship among the moonlighting proteins based on relative frequency of PP, NP, PP, and NN changes as obtained from polar, non-polar profiles. (B): Distance matrix consisting distances between pair of relative frequency (of PP, NP, PP, and NN changes) vectors corresponding to each moonlighting protein (distance less than 28 considered to be zero).

165 moonlighting protein sequences. A single instance of a poly-string length of 14 was observed in the moonlighting sequence P46060, composed of glutamic acid and occurring with a frequency of 1.

Additionally, poly-strings length of 10 appeared in sequences Q16637 and P19338, consisting of prolin and glutamic acid, respectively, each with a frequency of 1. Furthermore, P36897 contained a poly-string length of 9 comprised of alanine. Notably, sequences Q9P2J5 and O15105 exhibited poly-strings length of 7, featuring glutamic acid and glycine, respectively, both occurring once. Poly-strings lengths of 3, 4, 5, and 6 were identified in 130, 43, 18, and 9 many number of sequences, respectively. For a single sequence, poly-strings length of 7, 9, 10, and 14 were noticed with the highest frequency 1, while poly-string length of 3, 4, 5, 6 were noticed with the highest frequency 22, 8, 3, and 3, respectively. Noteworthy observations included three poly-strings of length 6 in P46060, all composed of glutamic acid, and three poly-strings of length 5 in both P49790 and P19338, originating from distinct amino acids, as detailed in Supplementary file-2. Furthermore, it was revealed that all 165 moonlighting proteins contained homogeneous poly-strings of lengths 1 and 2. Impressively, these were most prevalent in P49792, with frequencies of 2801 and 184, respectively. This observation emphasizes the common occurrence of short poly-strings composed of individual amino acids or consecutive pairs within these proteins.

The diversity observed in poly-string lengths for various amino acids

Table 7
Clusters of moonlighting proteins based on relative frequency of PP, NP, PP, and NN changes as obtained from Polar, non-polar profiles.

Clusters	Sequences
Cluster-1	{69,123,55,132,88,93,37,146,71,21,104,36,60,148,18,133,32, 125,31,155,3,8127,35,83,72,131,149,12,23,163,101,147,13, 22,70,84,90,110,17,157,81,47,45,52,108,61,66}
Cluster-2	{29,41,50,121,134}
Cluster-3	{5,86,38,46,89,114,107,111,128,137,162,16,96,80}
Cluster-4	{6,33,138,139,165,95,112,25,64,160,145,106,82,159,141,28,154, 49,65,118,140,9,43,109,115,59,136,99,135,39,117,51, 158,26,78,44,54,48,58,143,10,144,124,30,75,74,113,150,152,142}
Cluster-5	{42,116,156,67,98}
Cluster-6	{4,77,19,105,11,94,56,126,20,100,151}
Cluster-7	{24,40,73,122,63,85,76,91}
Cluster-8	{34,103,53,161,97,119,153,102,129,164}
Cluster-9	{2,7,57,68}
Cluster-10	{62,87}
Cluster-11	{79,92}

underscores the intricate arrangement and variability of amino acids within the moonlighting proteins, highlighting the complexity of their structural composition.

4.3. Polar, non-polar residue profiles of moonlighting proteins

The percentage distributions of polar and non-polar residues were computed for each moonlighting protein, as presented in Fig. 5. The analysis revealed that the average ratio of polar to non-polar residues was 1.0061 ± 0.2462. Notably, five sequences—Q9Y5Y2, Q99217, O00337, Q99714, and P19971—exhibited relatively lower ratios of 0.632, 0.63, 0.606, 0.601, and 0.58, respectively. On the higher end, the sequences Q9UQE7, Q5BJF6, P49759, P61254, P09429, and P62328 displayed elevated ratios of polar to non-polar residues, measuring 1.57, 1.62, 1.65, 1.79, 1.83, and 2.384, respectively. Furthermore, a distinct observation was made regarding the sequences P28482, P55263, P06744, and Q6YP21, where the percentages of polar and non-polar residues were exactly identical, resulting in a ratio of 1 for each sequence, as depicted in Fig. 5. This uniformity in the ratios underscores a specific compositional characteristic shared among these sequences.

The analysis delineates the distribution of polar and non-polar residues in moonlighting proteins, revealing a range of ratios that vary across different sequences. The observed variations provide insights into the potential functional and structural diversity inherent in these proteins.

4.3.1. Change response sequences of polar, non-polar profiles

Fig. 6 shows the relative frequency distribution of four changes namely PN, NP, PP, and NN. Percentages of polar to polar residue changes for the sequences P61254, P09429, P49759, and P62328 were significantly higher compared to others namely 40.28 %, 42.06 %, 42.65 %, and 51.16 %, respectively. Intriguingly, among these, P62328 displayed the lowest percentages of 'PN' (18.6 %), 'NN' (9.3 %), and 'NP' (20.9 %) changes. This emphasizes P62328's distinct residue change profile and suggests potential sequence-specific functional or structural implications. The observed variations in polar residue substitutions underscore the sequence-specific nature of residue changes and hint at their potential functional significance in these proteins.

Based on a distance threshold of 5, an analysis of 160 moonlighting sequences led to the formation of eleven distinct clusters, as depicted in the dendrogram (Fig. 7(A)) and summarized in Table 7. The most substantial cluster was comprised of 50 sequences, highlighted by a brown rectangular box in Fig. 7(A). Notably, sequence P62328 exhibited considerable dissimilarity from the majority of sequences. Particularly, P62328 displayed a distance greater than 28 from 131 other moonlighting sequences.

In the context of pairwise distances, five proteins—Q99714, P19971, Q9Y5Y2, O00337, and Q99217—were identified as the furthest from P62328. The calculated distances ranged from 45 to 49, with 49 representing the highest pairwise distance among all comparisons. These insights were visualized through a distance matrix, presented in Fig. 7 (B), with a defined threshold cut off at 28.

These findings underscore the diversity and relationships among the analyzed moonlighting sequences, revealing distinct clusters and the unique positioning of P62328 in relation to other sequences. The pairwise distances further highlight specific sequences that stand out in terms of dissimilarity from P62328, contributing to a comprehensive understanding of the sequence relationships within this dataset.

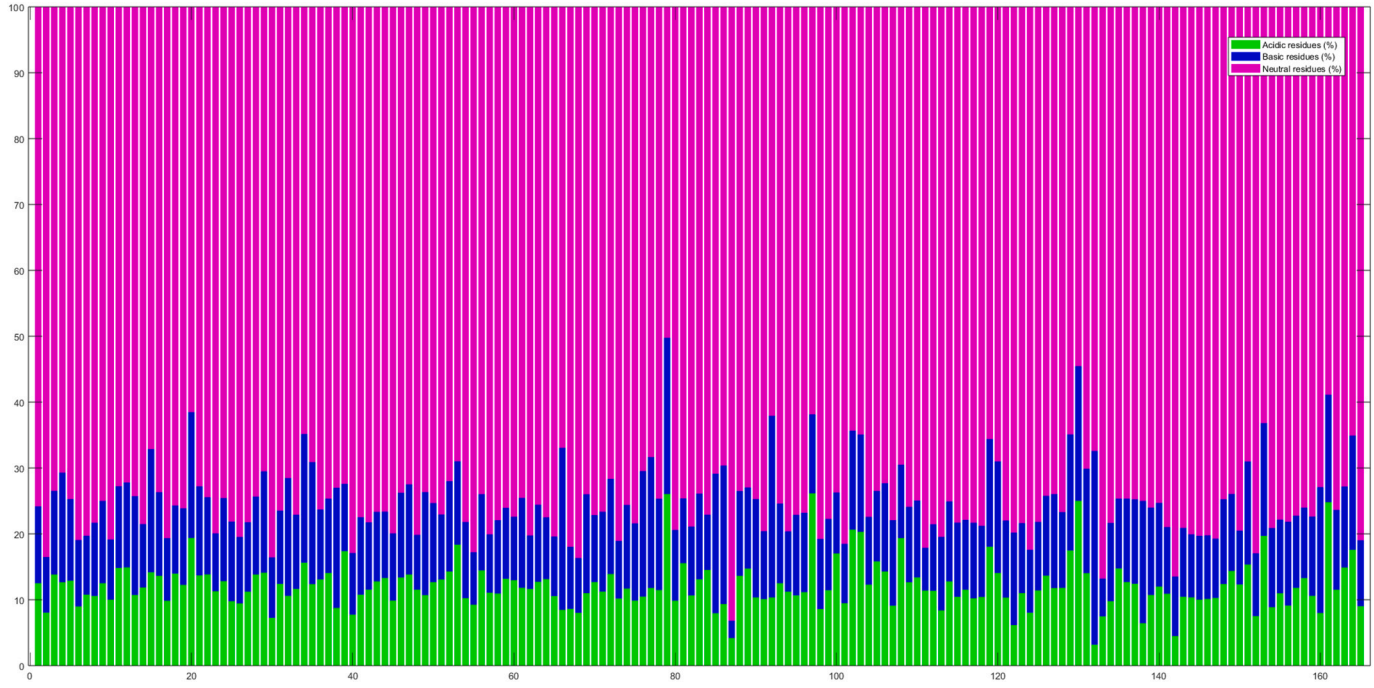


Fig. 8. Percentages of acidic, basic, neutral residues in moonlighting proteins.

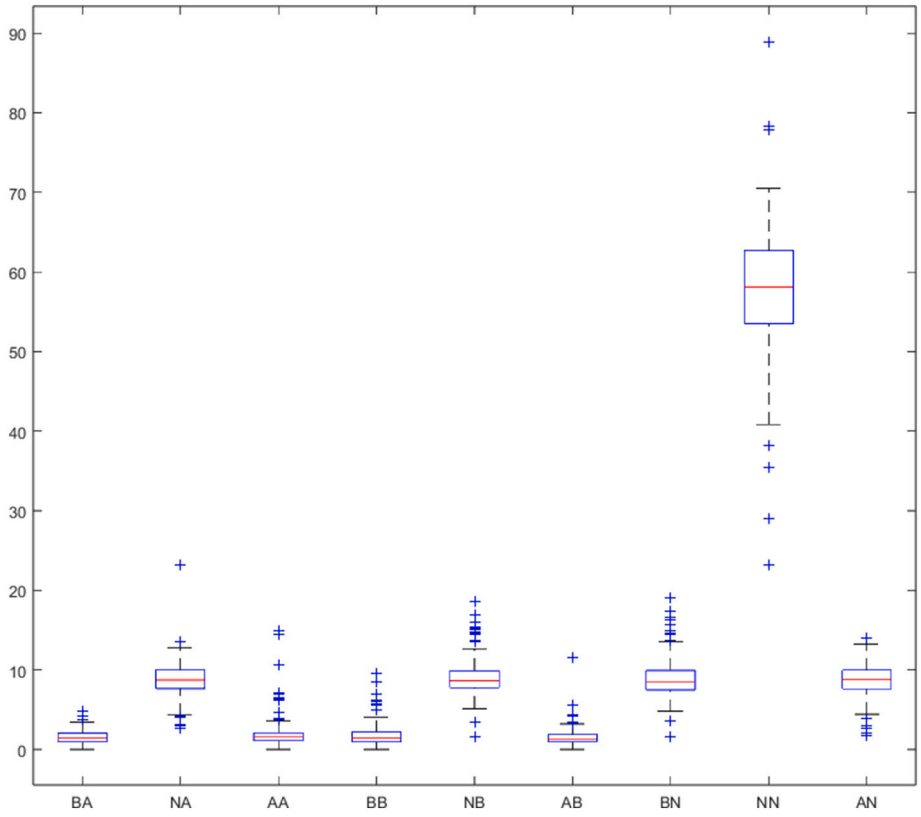


Fig. 9. Box-plot of the relative frequency of all nine changes in moonlighting proteins.

4.4. Acidic, basic, neutral residue based phylogenetic relationship

Based on the data illustrated in Fig. 8, the percentages of acidic, basic, and neutral residues were calculated for each moonlighting protein. In this context, the analysis encompassed 165 sequences, exhibiting varying percentages of neutral residues spanning from 50.2 % to 93.2 %. The calculated ratio of acidic to basic residue percentages was determined to be 1.03 ± 0.29, indicating a relatively balanced distribution between acidic and basic residues across the dataset. Furthermore, it was observed that the ratio of acidic to basic residues precisely equaled 1 for three specific sequences: P62826, P06733, and P18754.

The sequence P16402 stood out with distinct residue composition characteristics. It possessed the lowest percentage of acidic residues, measuring 3.17 %, and concurrently displayed the highest percentage of basic residues, at 29.41 %.

These findings collectively emphasize the diversity in residue compositions across the moonlighting proteins. The ratio of acidic and basic residues, as well as the particular characteristics of sequences such as P16402, provide insights into the potential functional implications of specific residue types within these proteins.

4.4.1. Change response sequences of acidic, basic, and neutral profiles

The relative frequency distribution of nine different residue changes, denoted as BA, NA, AA, BB, NB, AB, BN, NN, and AN, was visualized in Fig. 9. Among all the sequences, the highest percentages were observed for the BA (AB) changes, amounting to 4.74 % (11.63 %) in Q15311 (P62328). Notably, P62328 exhibited the highest percentages of NA (AN) changes, standing at 23.25 % (13.95 %), while P16402 had the lowest percentages, at 2.72 % (1.81 %) respectively.

In terms of specific changes, P09429 displayed the highest percentage of AA changes, accounting for 14.95 % of its residues. P16402, on the other hand, exhibited the highest percentage of BB changes, comprising 9.54 % of its residues. Interestingly, P16402 also stood out with the highest percentages of NB (BN) changes, reaching 18.64 % (19.1 %), whereas Q99217 had the lowest percentages, both recorded at 1.58 % (1.58 %). The highest percentage of NN changes was found in Q99217, constituting 88.95 % of its residues. Conversely, P62328 showcased the lowest percentage of NN changes, at 23.25 %.

These observations provide insights into the distinct residue change patterns across various sequences. The highest and lowest percentages of specific changes highlight the variability in these sequences and potentially suggest functional or structural significance associated with certain types of residue alterations.

Following change responses were absent in case of following moonlighting sequences:

- Basic to Acidic: Q9Y3E5
- Acidic to Acidic: P16402, P47992, Q7Z4W1, P09382, and P62328
- Basic to Basic: P09382
- Acidic to Basic: P09382, Q99988, and P62829

Under a distance threshold of 5, the analysis led to the formation of twelve distinct clusters comprising 144 moonlighting sequences. This information is visually depicted in the dendrogram (Fig. 10(A)) and summarized in Table 8. Notably, the largest cluster contained 56 sequences, indicated by a brown rectangular box in Fig. 10(A). Several sequences demonstrated considerable distance from others, reflecting their distinct positions within the dataset. Specifically, sequences P05455, P09429, Q99217, P62328, and P19338 were situated more than 30 distances away from 37, 117, 115, 150, and 32 sequences, respectively. The maximum calculated distance of 72.27 was observed between sequences P62328 and Q99217, underscoring the significant dissimilarity between these two.

These findings were effectively captured through a distance matrix, illustrated in Fig. 10(B), with a designated threshold cutoff set at 30.

Collectively, these results provide valuable insights into the structural relationships and variations among the analyzed moonlighting sequences. The identified clusters and pairwise distances contribute to a

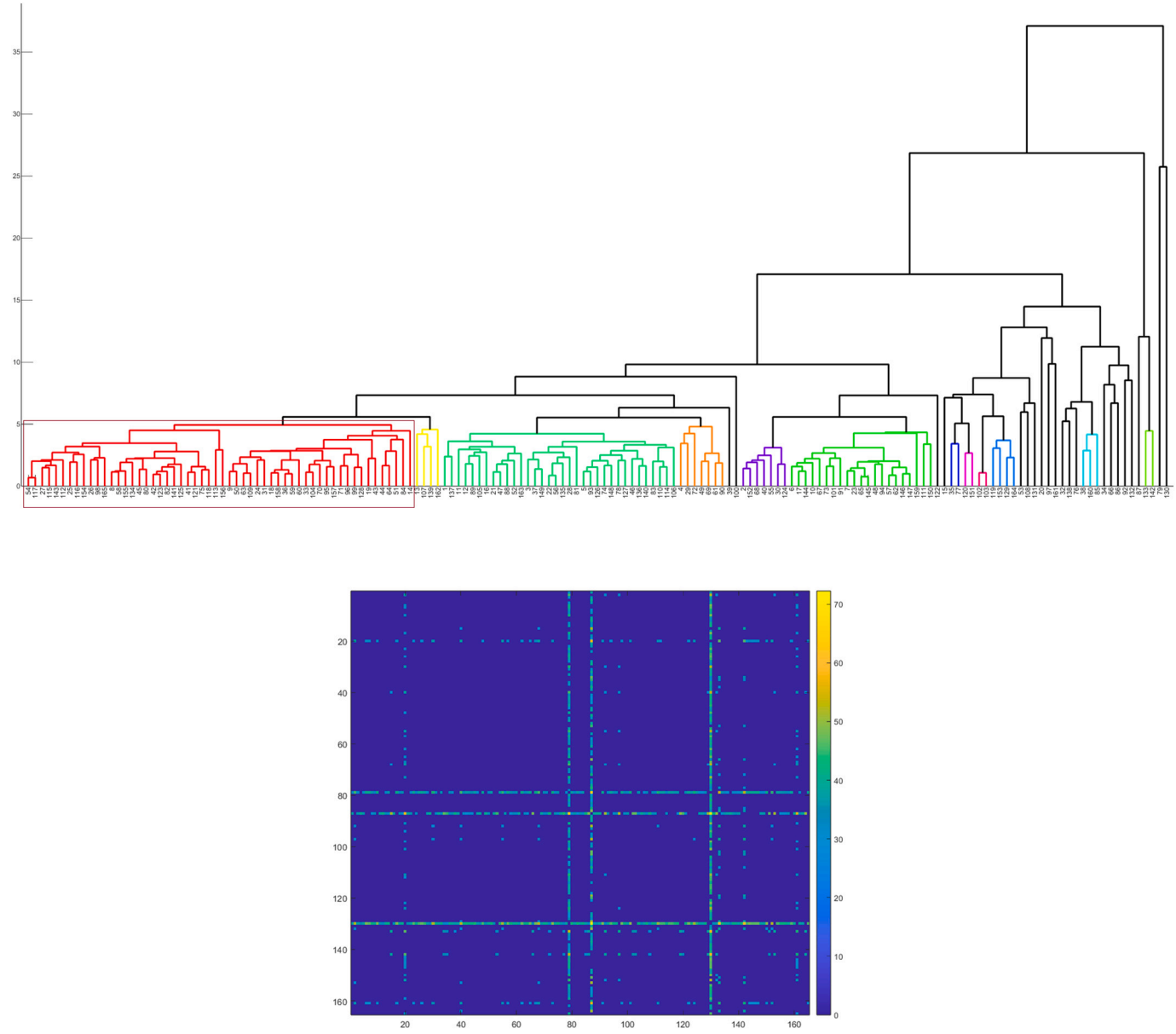


Fig. 10. (A): Phylogenetic relationship among the moonlighting proteins based on relative frequency of BA, NA, AA, BB, NB, AB, BN, NN, and AN changes as obtained from acidic, basic, and neutral profiles. (B): Distance matrix consisting distances between pair of relative frequency (of BA, NA, AA, BB, NB, AB, BN, NN, and AN changes) vectors corresponding to each moonlighting protein (distance less than 30 considered to be zero).

comprehensive understanding of the sequence associations within the dataset.

4.5. Intrinsic protein disorder analysis of moonlighting proteins

The intrinsic protein disorder probability of each residue was calculated for every moonlighting sequence. The percentages of disordered residues, highly flexible residues, moderately flexible residues, and other residues in each moonlighting protein were determined and presented in Fig. 11. It's worth noting that one sequence, P36969, was excluded from the intrinsic protein disorder analysis due to the presence of an ambiguous character 'U'.

An interesting observation was made regarding the moonlighting sequence P62328, which was found to consist entirely of disordered residues. Additionally, six moonlighting proteins exhibited a notably high percentage of disordered residues, namely P16402 (97.74 %), P49790 (94.64 %), Q96HA1 (92.79 %), O95997 (92.08 %), P09429 (91.16 %), and Q5BJF6 (88.66 %). Conversely, these proteins exhibited a significantly low percentage of highly flexible residues, with P16402 (2.26 %), Q96HA1 (3.28 %), P49790 (4.75 %), Q9Y6I3 (6.42 %), O95997 (7.92 %), P09429 (8.84 %), P19338 (10.42 %), and Q15311 (10.99 %).

In contrast, four sequences, namely Q86XF0 (63.64 %), P09382 (63.70 %), P05230 (67.10 %), and P07355 (67.85 %), displayed a notably high percentage of highly flexible residues. Remarkably, five moonlighting sequences, O95997, P09429, Q5BJF6, P62328, and P16402, did not exhibit any moderately flexible residues. The highest and lowest (non-zero) percentages of moderately flexible residues were observed in P15531 (54.61 %) and P49790 (0.61 %), respectively.

4.5.1. Global disorder status of human moonlighting proteins

Global disorder analysis revealed that human moonlighting proteins contain noticeable levels of intrinsic disorder. This is illustrated by Fig. 12A representing the results of the classification of the disorder

Table 8
Clusters of moonlighting proteins based on relative frequency of BA, NA, AA, BB, NB, AB, BN, NN, and AN changes as obtained from acidic, basic, and neutral profiles.

Cluster	Sequences
Cluster-1	{54,117,27,115,143,112,25,116,154,26,98,165,8,58,155, 134,45,80,42,123,82,141,125,41,121,75,118,113,156,9,50, 63,109,24,31,18,158,36,59,60,33,104,70,95,157,71,96,99, 128,19,43,44,64,51,84,14}
Cluster-2	{13,107,139,162}
Cluster-3	{1137,11,12,89,105,16,21,47,88,52,163,3,37,149,22,56, 135,28,81,5,93,126,74,148,78,127,46,136,140,83,110,114,106}
Cluster-4	{4,29,72,49,69,61,90}
Cluster-5	{2152,68,40,55,30,124}
Cluster-6	{6,17,144,10,67,73,101,91,7,23,65,145,48,94,57,62,146,147, 159,111,150}
Cluster-7	{35,77}
Cluster-8	{120,151}
Cluster-9	{102,103}
Cluster-10	{119,153,129,164}
Cluster-11	{38,160,85}
Cluster-12	{133,142}

status of these proteins based on the outputs of the per-residue disorder predictor PONDR® VSL2 in a form of the MDS vs. PPIDR dependence. This plot shows distribution of the moonlighting proteins within segments containing mostly ordered (blue and light blue), moderately disordered (pink and light pink), or mostly disordered (red) proteins. Based on these classifications, none of the moonlighting proteins was predicted as ordered by both MDS and PPIDR (dark blue segment is empty), and only 4 (2.4 %) of these were predicted as mostly ordered based on their MDS values (light blue segment). Therefore, remaining moonlighting proteins are expected to be either moderately or highly disordered. In fact, Fig. 12A shows that 50.3 % of these proteins were predicted as moderately disordered and containing noticeable levels of flexible or disordered residues/regions based on both their MDS and PPIDR values (dark pink segment). Additional 21.2 % proteins were expected to be moderately disordered based on their MDS values (light pink segment), whereas 26.1 % of the human moonlighting proteins are expected to be highly disordered (red segment). Importantly, most of these proteins (42 of 43) have PPIDR ≥50% and MDS ≥0.5.

Further support for relatively high disorder status of human moonlighting proteins was provided by combining the outputs of two binary predictors charge-hydropathy (CH) plot and cumulative distribution function (CDF) analysis that classify proteins as mostly ordered or mostly disordered (see Fig. 12B). The Result of the ΔCH-ΔCDF plot provides useful means for more detailed characterization of the global disorder status of proteins, classifying them as mostly ordered, molten globule-like or hybrid, or highly disordered (see Materials and Methods section). Fig. 12B shows that 67.3 % of the moonlighting proteins are located within the quadrant Q1 (bottom right corner) containing proteins predicted to be ordered by both predictors, whereas 32.7 % of these proteins are located outside the quadrant Q1 and can be considered as proteins with high disorder levels. Fig. 12B also shows that the quadrant Q2 (bottom left corner) corresponding to molten globular or hybrid proteins predicted to be ordered/compact by the CH-plot and disordered by the CDF analysis includes 18.2 % of these proteins. On the other hand, 9.7 % of moonlighting proteins, being located in the quadrant Q3 (top left corner), are predicted as highly disordered by both predictors, whereas the quadrant Q4 (top right corner) contains 4.8 % proteins predicted as disordered by CH-plot and ordered by CDF analysis.

4.5.2. Change response sequences of disordered, highly flexible, moderately flexible, and other residues profiles

In Fig. 13, the relative frequency distribution of sixteen changes derived from the change response sequence based on intrinsic disorder profiles is presented.

Notably, certain moonlighting protein sequences, specifically O00337, P02730, Q9H2X9, and P13569, exhibited significantly high percentages of O_O changes of residues, with values of 36.88 %, 22.42 %, 22.41 %, and 20.90 %, respectively. This suggests a notable propensity for these sequences to transition between other states.

An intriguing observation is that none of the 164 moonlighting protein sequences displayed changes from disordered to moderately flexible, disordered to other, highly flexible to other, moderately flexible

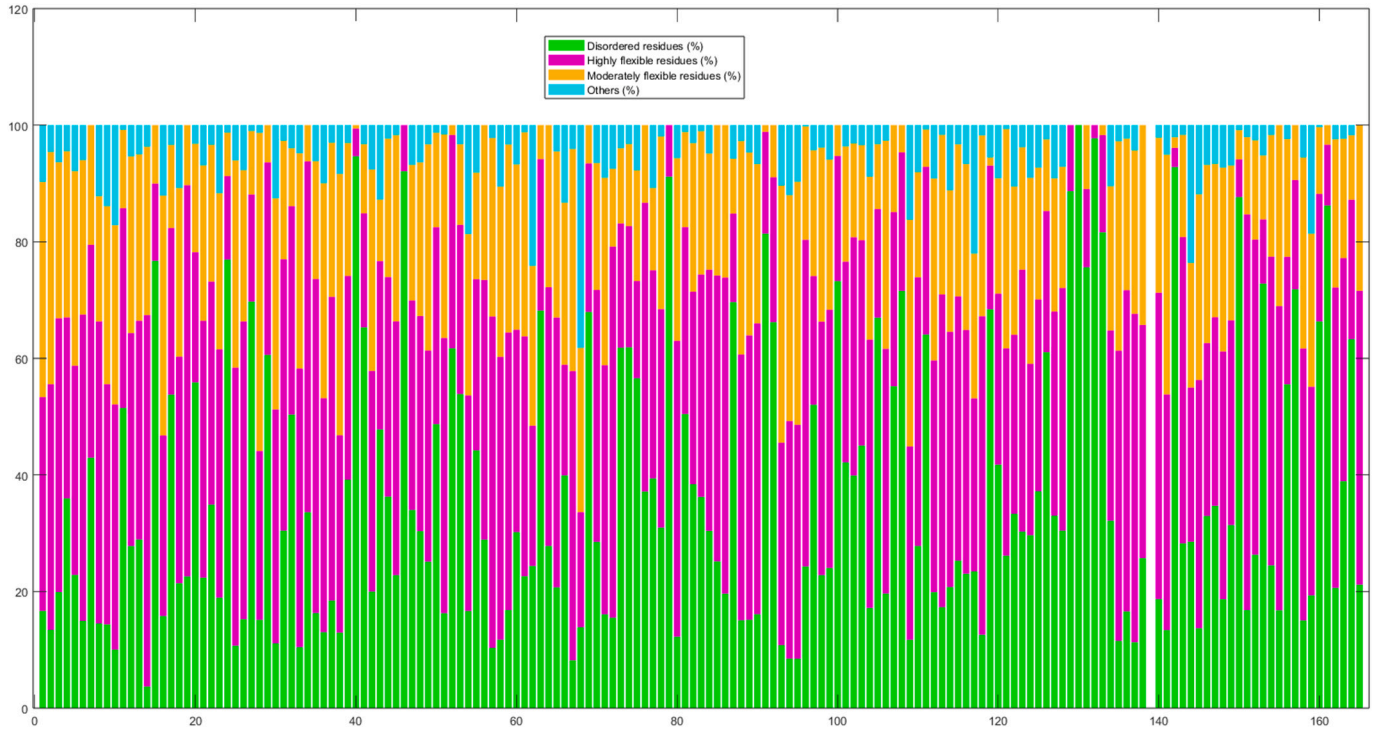


Fig. 11. Percentages of disordered, highly flexible, moderately flexible and other residues in moonlighting proteins.

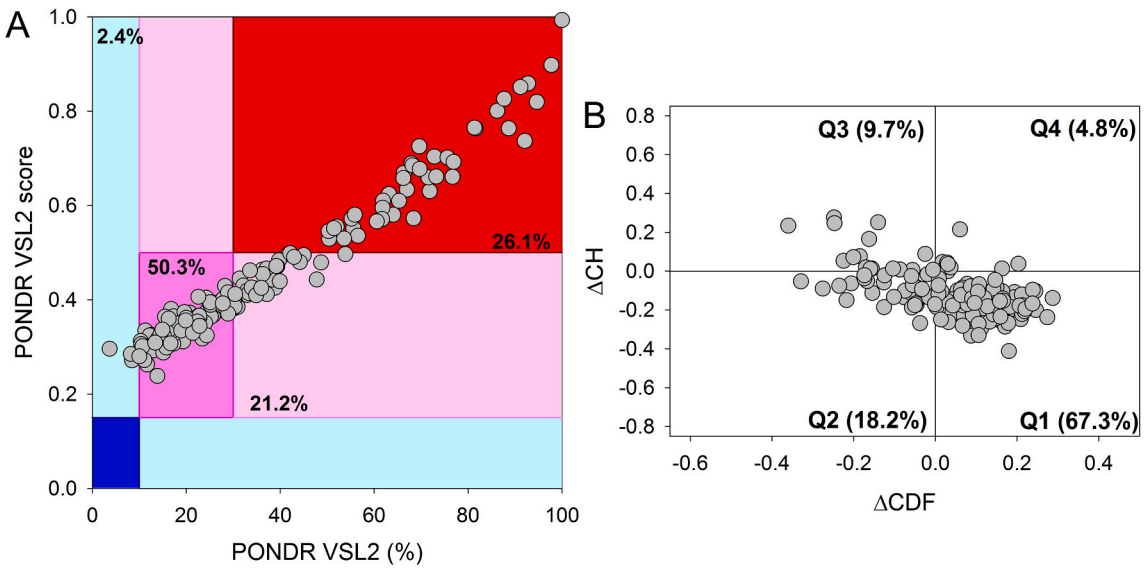


Fig. 12. Global disorder analysis of human moonlighting proteins. A. PONDR® VSL2 output for 165 human moonlighting proteins. PONDR® VSL2 score is the mean disorder score (MDS) for a protein, which is calculated as a sequence-length normalized sum of its per-residue scores. PONDR VSL2 (%) is a percent of predicted intrinsically disordered residues (PPIDR); i.e., percent of residues with disorder scores above 0.5. Color blocks indicate regions in which proteins are mostly ordered (blue and light blue), moderately disordered (pink and light pink), or mostly disordered (red). If the two parameters agree, the corresponding part of the background is dark (blue or pink), whereas light blue and light pink reflect areas in which only one of these criteria applies. The boundaries of the colored regions represent arbitrary and accepted cutoffs for MDS (y-axis) and the percentage of predicted disordered residues (PPDR; x-axis). Of the 165 human moonlighting proteins, 44 (26.6 %) are predicted to be highly disordered, 118 (71.5 %) demonstrate moderate disorder or conformational flexibility, and only 4 proteins (2.4 %) are predicted to be highly ordered. B. Charge-hydropathy and cumulative distribution function (CH-CDF) plot for 165 human moonlighting proteins. The Y-coordinate is calculated as the distance of the corresponding protein from the boundary in the CH plot. The X-coordinate is calculated as the average distance of the corresponding protein's CDF curve from the CDF boundary. The quadrant where the proteins are located determines their classification. Q1, 111 proteins (67.3 %) predicted to be ordered by CDF and compact/ordered by CH-plot. Q2, 30 proteins (18.2 %) predicted to be ordered/compact by CH-plot and disordered by CDF. Q3, 16 proteins (9.7 %) predicted to be disordered by CH-plot and CDF. Q4, 8 proteins (4.8 %) predicted to be disordered by CH-plot and ordered by CDF. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

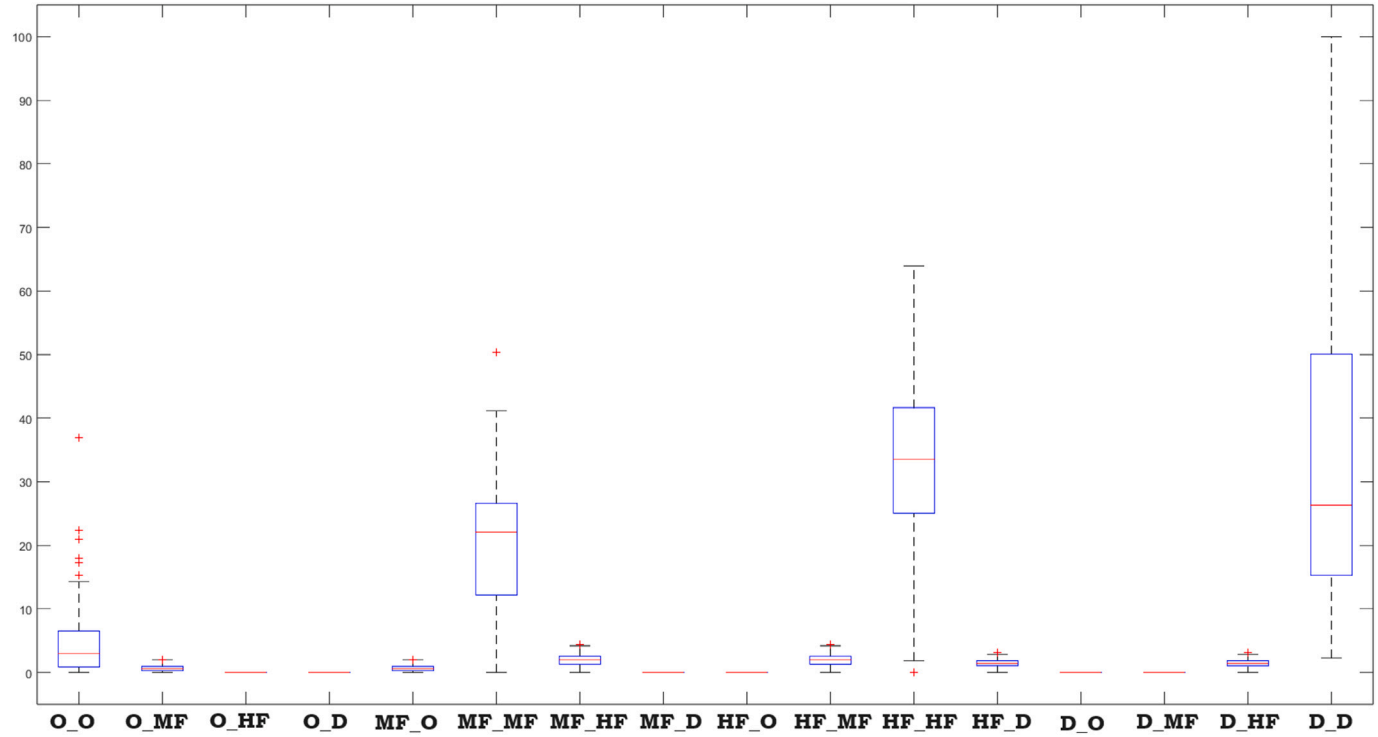


Fig. 13. Box-plot of the percentages of disordered, highly flexible, moderately flexible and other residues changes in moonlighting protein sequences.

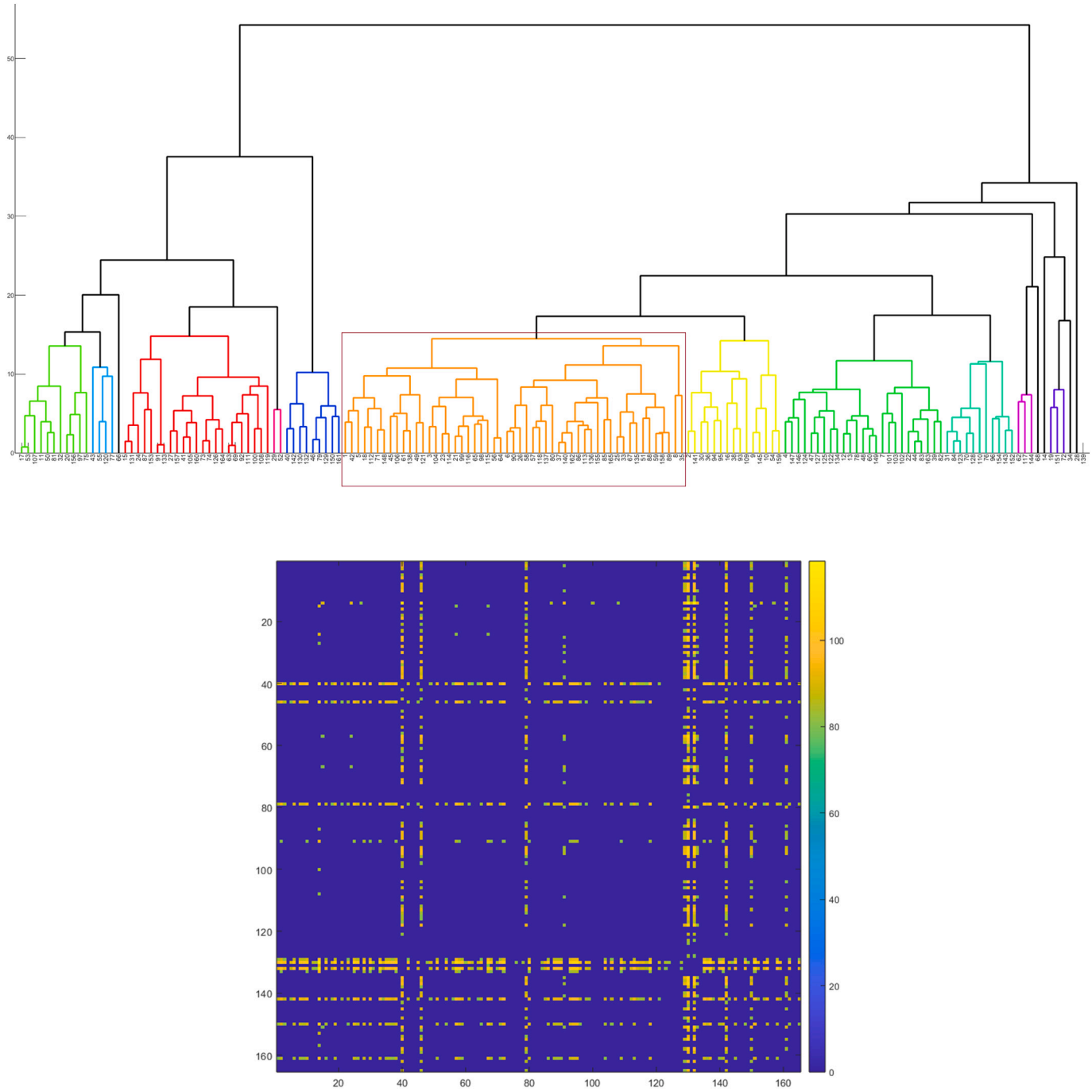


Fig. 14. (A): Phylogenetic relationship among the moonlighting proteins based on percentages of disordered, highly flexible, moderately flexible and other residues. (B): Distance matrix consisting distances between pair of percentages of disordered, highly flexible, moderately flexible and other residues corresponding to each moonlighting protein (distance less than 80 considered to be zero).

Table 9
Clusters of moonlighting proteins based on the change response profiles of disordered, highly flexible, moderately flexible, and other residues in moonlighting proteins.

Clusters	Sequences
Cluster-1	{17, 53, 107, 11, 50, 81, 32, 20, 156, 97, 75}
Cluster-2	{43, 55, 120, 77}
Cluster-3	{15, 131, 24, 87, 153, 91, 133, 27, 157, 41, 105, 160, 73, 74, 126, 164, 63, 69, 92, 111, 100, 108, 119}
Cluster-4	{29, 52}
Cluster-5	{40, 142, 130, 132, 46, 79, 129, 150, 161}
Cluster-6	{1, 42, 5, 18, 112, 71, 148, 45, 106, 61, 138, 49, 121, 3, 104, 23, 114, 21, 99, 116, 65, 98, 115, 56, 64, 6, 90, 26, 58, 57, 118, 137, 80, 37, 140, 162, 86, 113, 136, 155, 85, 165, 25, 33, 67, 135, 51, 88, 59, 158, 89, 8, 35}
Cluster-7	{2, 141, 30, 36, 94, 95, 16, 38, 93, 109, 9, 145, 10, 54, 159}
Cluster-8	{4, 147, 146, 124, 47, 127, 125, 122, 134, 12, 13, 78, 48, 60, 149, 7, 101, 103, 102, 22, 44, 83, 163, 39, 82}
Cluster-9	{31, 84, 123, 70, 128, 110, 76, 96, 154, 143, 152}
Cluster-10	{62, 117, 144}
Cluster-11	{19, 151, 72}

to disordered, other to disordered, or other to highly flexible residues. This lack of transitions between these specific states is worth noting. Furthermore, it was noticed that the highest amount (50.33 %) of changes from moderately flexible to moderately flexible (self-transition) was noticed in P15531 (Fig. 13). P07355 shows the highest percentage of (63.91 %) of changes from highly flexible to highly flexible residues. Fig. 13 provides valuable insights into the dynamic changes in intrinsic disorder profiles among the moonlighting protein sequences, highlighting specific sequences with distinctive disorder transition patterns.

An analysis of 159 moonlighting sequences, using a distance threshold of 15, resulted in the formation of eleven distinct clusters, which can be observed in the dendrogram (Fig. 14(A)) and are summarized in Table 9. The most substantial cluster, Cluster-6, consisted of 53 sequences and is highlighted by a brown rectangular box in Fig. 14 (A).

Interestingly, certain sequences, such as P49790, O95997, P09429, P62328, P16402, Q96HA1, Q9Y6I3, and P19338, were notably distant from a considerable number of other sequences, with distances greater than 80. Specifically, they were distant from 83, 73, 68, 98, 90, 80, 67, and 49 sequences, respectively. Among all possible pairwise distances, the maximum distance recorded was 118.34 for the pair P09382 and P62328. These insights were visually represented in a distance matrix, as shown in Fig. 7(B), with a defined threshold cutoff set at 80.

4.6. Phylogenetic relationship based on structural and physicochemical features

Under a distance threshold of 6.5, a comprehensive analysis revealed the formation of eleven distinct clusters encompassing 62 moonlighting sequences. This partitioning of sequences is visually represented in the dendrogram (Fig. 15(A)) and concisely detailed in Table 10. Notably, the most substantial cluster comprised 33 sequences, which is highlighted by a brown rectangular box in Fig. 15(A).

The analysis further revealed intriguing patterns of sequence distances. Specifically, sequences Q99217 and P62328 exhibited considerable dissimilarity from other sequences, as they were separated by a distance of more than 12, from 36 and 149 sequences, respectively. The maximum observed distance of 14.94 was recorded between sequences P07996 and P62328, underscoring a notable dissimilarity between these two sequences. These insights were concisely captured in a distance matrix, as illustrated in Fig. 15(B), with a predefined cutoff threshold of 12.

Collectively, the results shed light on the structural relationships and distinctions among the evaluated moonlighting sequences. The formation of clusters and the identification of specific sequences with notable dissimilarities contribute to an enhanced understanding of sequence associations within the dataset.

4.7. Cumulative clustering of moonlighting proteins

Utilizing a distance threshold of 150, an analysis of 156 moonlighting sequences resulted in the formation of six distinct clusters, as illustrated in the dendrogram (see Fig. 16(A)) and summarized in Table 11. The most extensive cluster, Cluster-3, encompassed 98 sequences and is highlighted within a brown rectangular box in Fig. 16(A).

The distance matrix highlighted certain sequences. Specifically, sequences P09429 (Serial No. 79), Q99217 (Serial No. 87), P62328 (Serial No. 130), and P16402 (Serial No. 132) were each positioned more than 250 units apart from 119, 120, 155, and 90 other sequences, respectively. A closer examination of the mutual distances between these four sequences showed that P09429 was situated within 250 units of both P62328 and P16402. Notably, the most substantial observed distance of 416.895 was observed between sequences Q99714 and P62328.

4.8. Agglomerated distal-proximal relationships of moonlighting proteins

We identified a total of 19 moonlighting proteins as outliers based on the analysis of various extracted features, as detailed in previous sections (see Table 12). Within this group of 19 moonlighting proteins, three proteins namely Q99217 (Amelogenin), P62328 (Thymosin beta-4), and P16402 (Histone H1.3) stood out as outliers (Table 12) in cumulative clustering also. These three proteins exhibited distinctive roles in diverse biological processes, showcasing their moonlighting functions:

Q99217 plays a significant role in maintaining mitochondrial DNA.

P62328 functions as a secreted chemotaxis ligand and regulates the activation of actin polymerization and ILK kinase.

P16402 serves as a receptor for thyroglobulin.

These 19 moonlighting proteins exhibited significant separation from the rest of the moonlighting proteins, as indicated by distinctive signature features originating from their amino acid sequences. It's important to note that this separation does not necessarily correlate with their distances based on moonlighting function, as moonlighting functionality is typically independent of the protein's primary structure [58].

Each of the six disjoint clusters as formed in cumulative clustering (Table 11) was further explored in detail to understand the proximal relationship among sequences of that cluster with respect to each feature individually. There may be more than one proximal set for a single feature and the sets were invariably disjoint. Proximal sets based on different features may or may not be disjoint. For example a subset of sequences belonging to cluster-1 formed one proximal set with respect to Feature-1 while producing three proximal sets with respect to Feature-5 (Table 13). Further, sequence numbers 79 and 161 of cluster-2 set up a proximal set only for feature 5 and not for any other feature. Thus Table 13 revealed how one or more than one proximal sets of moonlighting proteins with respect to individual feature finally contributed to the cumulative clustering. It is crucial to emphasize that the proximity of these sequential features does not necessarily imply a correlation with the moonlighting functions of the proteins. Following we present four types of examples of moonlighting proteins illustrating four possible different scenarios namely.

- **Case-I**: Shared no proximal relationship based on sequence-features, but having identical moonlighting functions.
- **Case-II**: Shared proximal relationship based on sequence-features, but having distinct moonlighting functions.
- **Case-III**: Shared proximal relationship based on sequence-features along with identical moonlighting functions.
- **Case-IV**: Shared no proximal relationship based on sequence-features along with distinct moonlighting functions.

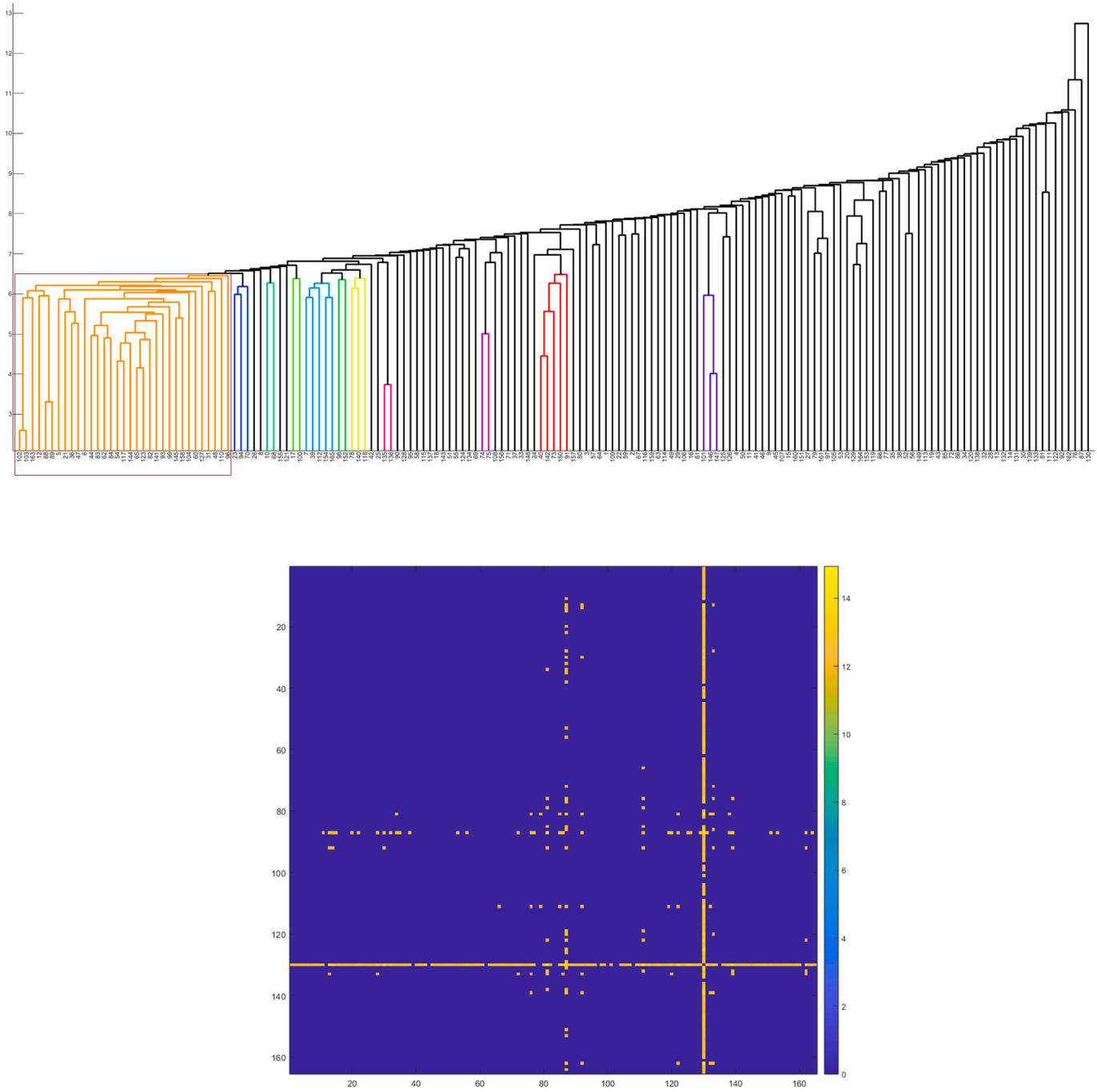


Fig. 15. (A): Phylogenetic relationship among the moonlighting proteins based on structural and physicochemical features. (B): Distance matrix consisting distances between pair of structural and physicochemical quantitative vectors corresponding to each moonlighting protein (distance less than 12 considered to be zero).

Table 10
Clusters of moonlighting proteins based on structural and physicochemical features.

Clusters	Sequences
Cluster-1	{102,103,163,12,88,89,5,21,36,47,6,44,83,62,84,54,117,144, 65,123, 82,141,93,99,145,158,104,60,127,31,48,110,96}
Cluster-2	{23,94,70}
Cluster-3	{10,68}
Cluster-4	{17,100}
Cluster-5	{7,39,112,154,165}
Cluster-6	{98,152}
Cluster-7	{78,140,118}
Cluster-8	{135,136}
Cluster-9	{74,75}
Cluster-10	{40,142,73,150,91}
Cluster-11	{101,146,147}

Example 1 (**Case-I**): Two proteins, P36897 (TGF-beta receptor type-1, Serial Number: 58) and Q99988 (Growth/differentiation factor 15, Serial Number: 107), are not proximal in any of the five sequence-features. P36897 and Q99988 belonged to cluster-3 and cluster-4 respectively in cumulative clustering (Table 11). What makes this particularly intriguing is that both of these proteins share an identical moonlighting function, namely, "Nucleus/Gene regulation."

Example 2 (**Case-II**): Two proteins, namely P08238 (Heat shock protein HSP 90-beta, Serial Number: 103) and P07900 (Heat shock protein HSP 90-alpha, Serial Number: 102), were observed to be in close proximity based on all five features. However, it's noteworthy that despite their sequence-based similarity, these two proteins have distinct moonlighting functions. Specifically, HSP 90-alpha serves as a chaperone, contributing to MMP2 activation and wound healing through its involvement in exosomes, while HSP 90-beta plays a pivotal role in regulating the transcription machinery.

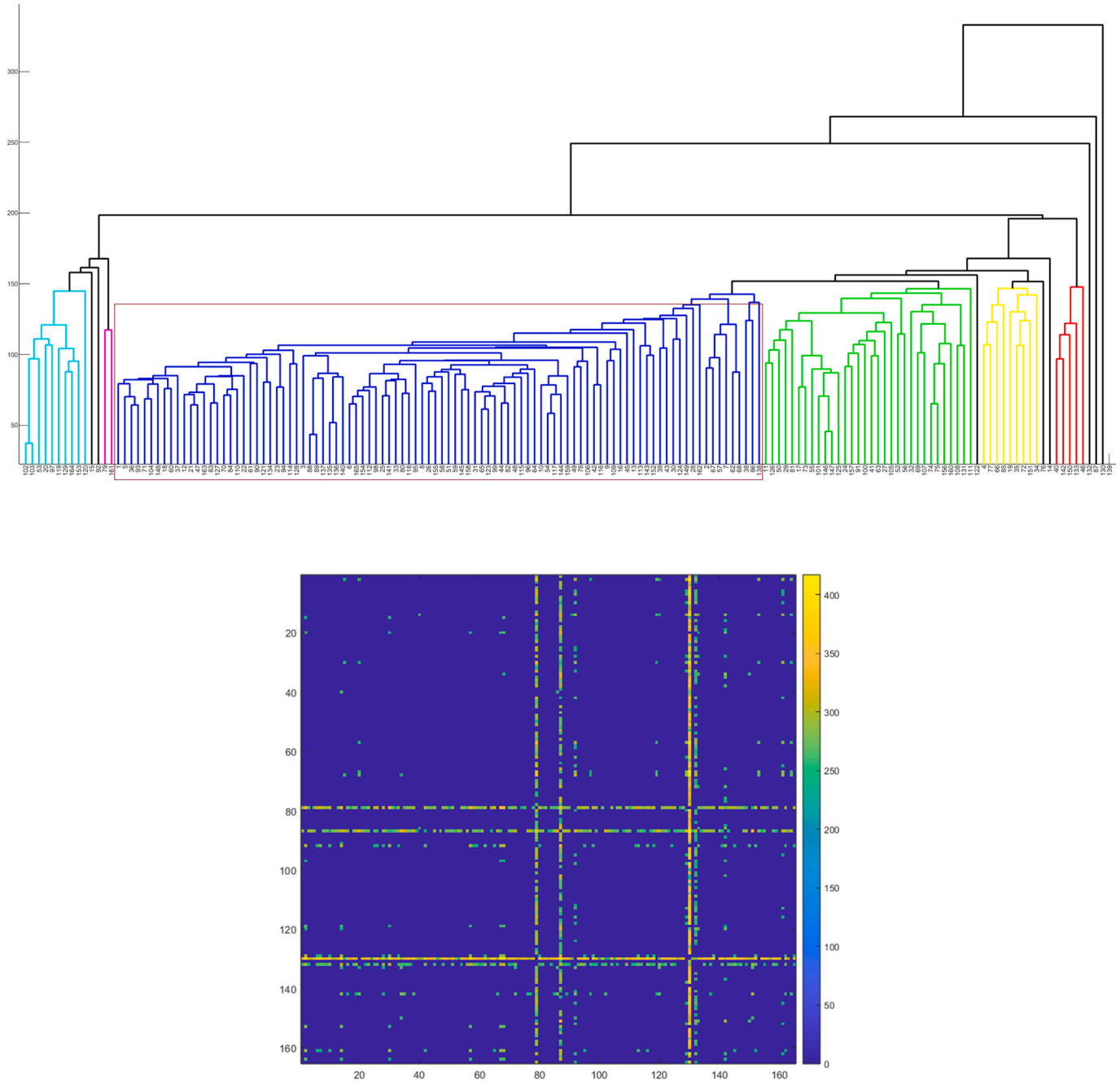


Fig. 16. (A): Phylogenetic relationship among the moonlighting proteins from cumulative clustering. (B): Corresponding distance matrix (distance less than 250 considered to be zero).

Table 11
Cumulative clustering of moonlighting proteins based on five features.

Clusters	Sequences
Cluster-1	{102, 103, 53, 20, 97, 119, 129, 164, 153, 120}
Cluster-2	{79, 161}
Cluster-3	{1, 5, 36, 93, 71, 104, 148, 18, 60, 37, 12, 21, 47, 163, 83, 127, 70, 84, 110, 22, 61, 90, 121, 134, 23, 94, 114, 128, 3, 88, 89, 137, 135, 136, 140}
Cluster-3	{6, 165, 154, 112, 98, 25, 141, 33, 80, 118, 95, 8, 26, 155, 58, 51, 59, 145, 158, 31, 65, 123, 99, 44, 82, 48, 115, 96, 64, 10, 54, 117, 144, 159, 49, 78}
Cluster-3	{106, 42, 116, 9, 109, 16, 45, 13, 113, 143, 152, 39, 43, 30, 124, 149, 28, 162, 2, 67, 57, 7, 62, 68, 38, 86, 138}
Cluster-4	{11, 126, 50, 29, 81, 17, 73, 55, 101, 146, 147, 125, 24, 157, 91, 100, 41, 63, 27, 105, 52, 56, 32, 69, 107, 74, 75, 156, 160, 108, 131, 111}
Cluster-5	{4, 77, 66, 85, 19, 35, 72, 151, 34}
Cluster-6	{40, 142, 150, 133, 46}

Example 3 (**Case-III**): Our sequence-based phylogenetic analysis revealed that two moonlighting proteins, P02511 (alpha-crystallin B chain, Serial Number: 52) and P02489 (alpha-crystallin A chain, Serial Number: 56), share the same cluster (Cluster-4, as shown in Table 11). These proteins exhibit identical moonlighting functions as "Heat-shock proteins."

Example 4 (**Case-IV**): Q99217 (Amelogenin, Serial Number: 87) and P62328 (Thymosin Beta-4 (sequester actin), Serial Number: 130) turned out to be distal based on cumulative sequential features, and these two proteins have different moonlighting functions (Table 14).

4.9. Discussion

Moonlighting proteins, distinguished by their multifunctional nature, hold a significant role in shaping the trajectory of evolution. Beyond their biochemical versatility, these proteins emerge as crucial hotspots, influencing the adaptive strategies of organisms in response to environmental changes. Their ability to engage in multiple functions facilitates accelerated evolutionary responses, allowing organisms to swiftly adapt to new challenges. Furthermore, moonlighting proteins contribute to evolutionary innovation by serving as platforms for the emergence of novel functions within a single polypeptide chain. This innovation, coupled with their role in genomic economy, underscores their importance as dynamic elements in the evolutionary landscape. Additionally, moonlighting proteins act as fine-tuned regulators of biological processes, providing nuanced adjustments in response to selective pressures. Their multifunctionality not only conserves genomic resources but also plays a pivotal role as checkpoints in the ongoing story of Darwinian selection. In essence, moonlighting proteins stand as versatile actors, driving innovation, and embodying the intricate interplay of adaptation and evolution. One-hundred and sixty-five human moonlighting proteins have been obtained from the database MultitaskProtDBII, spanning a wide array of types, including receptors, enzymes, transcription factors, adhesins, and scaffolds. Remarkably, these proteins exhibit diverse combinations of canonical functions. While some of these proteins can execute both functions simultaneously, others adapt their role in response to environmental changes. The wide-ranging examples of identified moonlighting proteins and their potential benefits to organisms, such as orchestrating cellular activities, strongly suggest that many more moonlighting proteins remain to be unveiled. Furthermore, it was reported in the literature that amino acid sequence homology of moonlighting proteins may not necessarily possess moonlighting functions themselves [2,8]. Therefore, relying solely on sequence homology might not be adequate for accurately predicting protein function. In this work, we investigate 165 human moonlighting proteins to discover various signature-features, which led us to develop a phylogenetic relationship among the set of moonlighting proteins. Signature-features include frequency of changes as obtained from change response profile based on three features, namely polar-nonpolar, acidic-basic-neutral, and intrinsic protein disorder. Phylogenetic analyses revealed a set of 19 moonlighting proteins having a distal relationship from remaining moonlighting proteins as obtained from different features (Table 12)). In the cumulative clustering, 156 moonlighting proteins formed six different clusters (Table 11) which were

Table 12
List of moonlighting proteins which turned out to be outliered based on various features namely Feature-1 : Relative frequency of amino acids; Feature-2 : Homogeneous poly-string frequency; Feature-3 : Relative frequency of four changes based on polar and nonpolar residues; Feature-4 : Relative frequency of nine changes based on acidic, basic and neutral residues Feature-5 : Relative frequency of sixteen changes based on disordered, highly flexible, moderately flexible, and other residues; Feature-6 : Structural and physicochemical features.

Serial Number	Uniprot ID	With respect to
21	Q9P2J5	Feature-2
27	Q16637	Feature-2
39	P46060	Feature-2
40	P49790	Feature-2 , Feature-5
46	O95997	Feature-5
58	P36897	Feature-2
75	O15105	Feature-2
79	P09429	Feature-3 , Feature-4 , Feature-5
87	Q99217	Feature-1 , Feature-4 , Feature-6 , Cumulative
92	P61254	Feature-3
120	P49759	Feature-3
126	P49792	Feature-2
129	Q5BJF6	Feature-5
130	P62328	Feature-3 , Feature-4 , Feature-5 , Feature-6 , Cumulative
132	P16402	Feature-1 , Feature-5 , Cumulative
133	P37198	Feature-4
142	Q96HA1	Feature-4 , Feature-5
150	Q9Y6I3	Feature-5
161	P19338	Feature-2 , Feature-5

Table 13
Proximal sets of moonlighting proteins.

Cluster-1: {102, 103, 53, 20, 97, 119, 129, 164, 153, 120}	
Serial numbers	***With respect to***
{102, 103, 53, 20, 119, 129, 164, 153}	Feature-1
{102, 103}, {53, 20, 97, 120}, {119, 164, 153}	Feature-5
{102, 103, 129, 164, 153, 120}	Feature-4
{102, 103}	Feature-6
{102, 103, 53, 97, 119, 129, 164, 153}	Feature-3
Cluster-2: {79, 161}	
Serial numbers	***With respect to***
{79, 161}	Feature-5
Cluster-3: {1, 5, 36, 93, 71, 104, 148, 18, 60, 37, 12, 21, 47, 163, 83, 127, 70, 84, 110, 22, 61, 90, 121, 134, 23, 94, 114, 128, 3, 88, 89, 137, 135, 136, 140, 6, 165, 154, 112, 98, 25, 141, 33, 80, 118, 95, 8, 26, 155, 58, 51, 59, 145, 158, 31, 65, 123, 99, 44, 82, 48, 115, 96, 64, 10, 54, 117, 144, 159, 49, 78, 106, 42, 116, 9, 109, 16, 45, 13, 113, 143, 152, 39, 43, 30, 124, 149, 28, 162, 2, 67, 57, 7, 62, 68, 38, 86, 138}	
Serial numbers	***With respect to***
{1, 5, 36, 93, 71, 104, 148, 18, 60, 37, 12, 21, 47, 163, 83, 127, 70, 84, 110, 22, 61, 90, 121, 134, 23, 94, 114, 128, 3, 88, 89, 141, 8, 26, 155, 58, 51, 59, 145, 158, 65, 123, 99, 44, 82, 48, 96, 64, 117, 144, 159, 109, 16, 124, 28}	Feature-1
{137, 135, 136, 140, 6, 165, 154, 112, 98, 25, 33, 80, 118, 95, 31, 115, 54, 78, 152}	Feature-1
{49, 106, 9}	Feature-1
{42, 116, 57, 7, 62, 68}	Feature-1
{2, 67}	Feature-1
{1, 5, 93, 148, 37, 12, 21, 47, 163, 83, 127, 110 22 61 90, 114, 3, 88, 89, 137, 135, 136, 140, 49 78 106, 16, 149, 28}	Feature-4
{36, 71, 104, 18, 60, 70, 84, 121, 134, 128, 165, 154, 112, 98, 25, 141, 33, 80, 118,95, 8, 26, 155, 58, 51, 59, 158, 31, 123, 99, 44, 82, 115, 96, 64, 54, 117, 42, 116, 9, 109, 45, 13, 113, 143, 43, 162}	Feature-4
{23, 94, 6, 145, 65, 48, 10, 144, 159, 152, 30, 124, 2, 67, 57, 7, 62, 68}	Feature-4
{1, 5, 71, 104, 148, 18, 37, 21, 61, 90, 121, 23, 114, 3, 88, 89, 137, 135, 136, 140, 6, 165, 112, 98, 25, 33, 80, 118, 8, 26, 155, 58, 51, 59, 158, 65, 99, 115, 64, 49, 106, 42, 116, 45, 113, 162, 67, 57, 86, 138}	Feature-5
{36, 93, 94, 141, 95, 145, 10, 54, 159, 9, 109, 16, 30, 2, 38}	Feature-5
{60, 12, 47, 163, 83, 127, 22, 134, 44, 82, 13, 39, 124, 149, 7}	Feature-5
{70, 84, 110, 128, 154, 31, 123, 96, 143, 152}	Feature-5
{117, 144, 62}	Feature-5
{5, 36, 93, 104 60, 12, 21, 47, 163, 87, 127, 84, 110, 89, 141, 145, 158, 31, 65, 123, 99, 44, 82, 48, 96, 54, 117, 144, 62}	Feature-6
{70, 23 94}	Feature-6
{135 136}	Feature-6
{140, 118, 78}	Feature-6
{98, 151}	Feature-6
{165, 154, 112, 39, 7}	Feature-6
{10, 68}	Feature-6
{5, 114, 128, 89, 137, 80, 96, 16, 162, 38, 86}	Feature-3
{36, 93, 71, 104, 148, 18, 60, 37, 12, 21, 47, 163, 83, 127, 70, 84, 110, 22, 61, 90, 23, 3, 88, 8, 155, 31, 123, 45, 13, 149}	Feature-3
{135, 136, 140, 6, 165, 154, 112, 25, 141, 33, 118, 95, 26, 58, 51, 59, 145, 158, 65, 99, 44, 82, 48, 115, 96, 64, 10, 54, 117, 144, 159, 49, 78, 106, 9, 109, 113, 143, 152, 39, 43, 30, 124, 28, 138}	Feature-3
{98, 42, 116, 67}	Feature-3
{2, 57, 7, 68}	Feature-3

Table 14
Proximal sets of moonlighting proteins.

Cluster-4: {11, 126, 50, 29, 81, 17, 73, 55, 101, 146, 147, 125, 24, 157, 91, 100, 41, 63, 27, 105, 52, 56, 32, 69, 107, 74, 75, 156, 160, 108, 131, 111}	
Serial numbers	***With respect to***
{11, 126, 24, 100}, {81, 111}, {17, 73, 55, 101, 146, 147, 125, 41, 63}, {157, 91, 105}	Feature-1
{52, 56}, {69, 107}, {74, 75, 156, 160, 108, 131}	Feature-1
{11, 126, 81, 105, 52, 56, 74}, {29, 69}	Feature-4
{17, 73, 101, 146, 147, 91, 111}, {125, 24, 157, 41, 63, 27, 75, 156}	Feature-4
{11, 50, 81, 17, 55, 32, 107, 75, 156}, {29, 52}, {101, 146, 147, 125},	Feature-5
{126, 73, 24, 157, 91, 100, 41, 63, 27, 105, 69, 74, 160, 108, 131, 111}	Feature-5
{17, 100}, {73, 91}, {101, 146, 147}, {74, 75, 108}	Feature-6
{11, 126, 100, 105, 56}, {73, 24, 91, 63}, {107, 111}	Feature-3
{50, 29, 81, 17, 55, 101, 146, 147, 125, 157, 41, 52, 32, 69, 74, 75, 108, 131}	Feature-3
Cluster-5: {4, 77, 66, 85, 19, 35, 72, 151, 34}	
Serial numbers	***With respect to***
{77, 66, 35}	Feature-1
{4, 72}, {35, 77, 151}, {66, 34}	Feature-4
{85, 35}, {19, 72, 151, 34}	Feature-5
{4, 77, 19, 151}, {66, 35, 72}	Feature-3
Cluster-6: {40, 142, 150, 133, 46}	
Serial numbers	***With respect to***
{40, 150}, {142, 131}	Feature-4
{40, 142, 150, 46}	Feature-5
{40, 142, 150}	Feature-6
{142, 150}	Feature-3

further analyzed in terms of proximity among moonlighting proteins for each sequence based feature. This distal-proximal relationship among these moonlighting proteins draws a distinct characterization of human moonlighting proteins, using unique genomic imprints that define these proteins within the human proteome. Furthermore, distal-proximal relationships might provide insights into the divergence and convergence of moonlighting protein functions.

While our study offers valuable insights into the diversity and phylogenetic relationships of 165 human moonlighting proteins, it is important to acknowledge certain limitations. The dataset primarily relies on the information available in the MultitaskProtDBII, and potential biases in this database could influence the comprehensiveness of our analysis. Additionally, the complexity of moonlighting protein functions and their adaptive responses to environmental changes may not be fully captured by the selected sequence-based features, indicating the need for complementary experimental validations to further refine our understanding of these multifaceted proteins.

This investigation sheds light on unique systems biology properties of the proteins and their association with certain physiological pathways especially for the characterization of the human multi-domain proteins. The study uncovers intriguing questions and hypotheses regarding the functional diversity of moonlighting proteins. These could serve as the basis for future research endeavors in the fields of moonlighting proteomics.

The future scope of this study involves expanding the dataset of human moonlighting proteins, utilizing additional data sources and advanced bioinformatics tools for a more comprehensive analysis. Integration of predictive models incorporating diverse features beyond sequence-based characteristics, coupled with experimental validations through functional assays and structural studies, will enhance the accuracy of our predictions. Investigating the functional implications of the identified distal-proximal relationships among moonlighting proteins provides an avenue for understanding the evolutionary dynamics shaping their multifunctional nature. Exploring the unique genomic imprints identified in this study may uncover additional regulatory elements governing moonlighting proteins, necessitating collaborative efforts with experimental biologists for a more holistic understanding of their roles in adaptation and evolution.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijbiomac.2023.128998.

Author statement

The manuscript was prepared without using any generative artificial intelligence (AI) or AI-assisted technologies in the writing process.

The manuscript was prepared without experiments with humans or animals.

The authors declare that all authors have seen and approved the final version of the manuscript being submitted.

The authors warrant that the article is the authors' original work, has not received prior publication, and is not under consideration for publication elsewhere.

CRediT authorship contribution statement

Debaleena Nawn: Formal analysis, Investigation, Methodology, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Sk. Sarif Hassan:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Moumita Sil:** Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Ankita Ghosh:** Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Arunava Goswami:** Conceptualization, Formal analysis, Investigation, Methodology,

Visualization, Writing – original draft, Writing – review & editing. **Pallab Basu:** Formal analysis, Investigation, Methodology, Writing – review & editing. **Guy W. Dayhoff:** Formal analysis, Investigation, Methodology, Writing – review & editing. **Kenneth Lundstrom:** Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Vladimir N. Uversky:** Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

Authors are very grateful to the laboratory supporters Mr. Ashim Dhar and Mrs. Piyali Maitra for their logistic assistance. Especially, authors would like to thank wholeheartedly Mr. Arindam Samanta for his insightful comments. Authors would like to thank the Department of Biotechnology, Govt. of India (BT/PR26647/NNT/28/1365/2017), Indian Space Research Organisation (ISRO/RES/3/827/19-20), and Indian Statistical Institute (ISI) (ISI/TAC/PROJECT-1/2023-24) for their financial support.

References

[1] D.H. Huberts, I.J. van der Klei, Moonlighting proteins: an intriguing mode of multitasking, Biochim. Biophys. Acta, Mol. Cell Res. 1803 (4) (2010) 520–525.
[2] C.J. Jeffery, Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins, Curr. Opin. Struct. Biol. 14 (6) (2004) 663–668.
[3] J. Piatigorsky, W.E. O'Brien, B.L. Norman, K. Kalumuck, G.J. Wistow, T. Borras, J. M. Nickerson, E.F. Wawrousek, Gene sharing by delta-crystallin and argininosuccinate lyase, Proc. Natl. Acad. Sci. 85 (10) (1988) 3479–3483.
[4] C.J. Jeffery, Moonlighting proteins, Trends Biochem. Sci. 24 (1) (1999) 8–11.
[5] C.J. Jeffery, Multifunctional proteins: examples of gene sharing, Ann. Med. 35 (1) (2003) 28–35.
[6] J. Piatigorsky, G.J. Wistow, Enzyme/crystallins: gene sharing as an evolutionary strategy, Cell 57 (2) (1989) 197–199.
[7] A.W. Torrance, Gene concepts, gene talk, and gene patents, Minn. JL Sci. & Tech. 11 (2010) 157.
[8] C.J. Jeffery, Moonlighting proteins—an update, Mol. BioSyst. 5 (4) (2009) 345–350.
[9] M.A. Fares, The evolution of protein moonlighting: adaptive traps and promiscuity in the chaperonins, Biochem. Soc. Trans. 42 (6) (2014) 1709–1714.
[10] B. Henderson, M.A. Fares, A.C. Martin, Protein Moonlighting in Biology and Medicine, John Wiley & Sons, 2016.
[11] S.D. Copley, An evolutionary perspective on protein moonlighting, Biochem. Soc. Trans. 42 (6) (2014) 1684–1691.
[12] N. Singh, N. Bhalla, Moonlighting proteins, Annu. Rev. Genet. 54 (2020) 265–285.
[13] C.J. Jeffery, Why Study Moonlighting Proteins?, 2015.
[14] A. Espinosa-Cantú, D. Ascencio, F. Barona-Gómez, A. DeLuna, Gene duplication and the evolution of moonlighting proteins, Front. Genet. 6 (2015) 227.
[15] M.N. Gupta, A. Alam, S.E. Hasnain, Protein promiscuity in drug discovery, drug-repurposing and antibiotic resistance, Biochimie 175 (2020) 50–57.
[16] M.N. Gupta, V.N. Uversky, Moonlighting enzymes: when cellular context defines specificity, Cell. Mol. Life Sci. 80 (5) (2023) 130.
[17] C.J. Jeffery, Proteins with neomorphic moonlighting functions in disease, IUBMB Life 63 (7) (2011) 489–494.
[18] G. Sriram, J.A. Martinez, E.R. McCabe, J.C. Liao, K.M. Dipple, Single-gene disorders: what role could moonlighting enzymes play? Am. J. Hum. Genet. 76 (6) (2005) 911–924.
[19] J. Ovádi, Moonlighting Proteins in Neurological Disorders, 2011.
[20] Q.-H. Vuong, A New Theory of Serendipity: Nature, Emergence and Mechanism, Walter De Gruyter GmbH, 2022.
[21] S. Hernández, L. Franco, A. Calvo, G. Ferragut, A. Hermoso, I. Amela, A. Gómez, E. Querol, J. Cedano, Bioinformatics and moonlighting proteins, Front. Bioeng. Biotechnol. 3 (2015) 90.
[22] V. Vacic, V.N. Uversky, A.K. Dunker, S. Lonardi, Composition profiler: a tool for discovery and visualization of amino acid composition differences, BMC 8 (2007) 1–7.
[23] S.S. Hassan, M. Sil, S. Chakraborty, A. Goswami, P. Basu, D. Nawn, V.N. Uversky, Possible functional proximity of various organisms based on the bioinformatics analysis of their taste receptors, Int. J. Biol. Macromol. 222 (2022) 2105–2121.
[24] C. Nacar, The frequencies of amino acids in secondary structural elements of globular proteins, Clin. Exp. Health Sci. 13 (2) (2023) 84–89.
[25] S. Hormoz, Amino acid composition of proteins reduces deleterious impact of mutations, Sci. Rep. 3 (1) (2013) 2919.
[26] B.J. Strait, T.G. Dewey, The shannon information entropy of protein sequences, Biophys. J. 71 (1) (1996) 148–155.
[27] S. Shen, B. Kai, J. Ruan, J.T. Huzil, E. Carpenter, J.A. Tuszynski, Probabilistic analysis of the frequencies of amino acid pairs within characterized protein sequences, Physica A 370 (2) (2006) 651–662.
[28] M.Y. Karpeisky, V. Ilyin, Analysis of non-polar regions in proteins, J. Mol. Biol. 224 (3) (1992) 629–638.
[29] H. Xiong, B.L. Buckwalter, H.-M. Shieh, M.H. Hecht, Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides, Proc. Natl. Acad. Sci. 92 (14) (1995) 6349–6353.
[30] S. Kamtekar, J.M. Schiffer, H. Xiong, J.M. Babik, M.H. Hecht, Protein design by binary patterning of polar and nonpolar amino acids, Science 262 (5140) (1993) 1680–1685.
[31] B. Mészáros, G. Erdős, Z. Dosztányi, Iupred2a: context-dependent prediction of protein disorder as a function of redox state and protein binding, Nucleic Acids Res. 46 (W1) (2018) W329–W337.
[32] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, A.K. Dunker, Exploiting heterogeneous sequence properties improves prediction of protein disorder, Proteins: Struct., Funct., Bioinf. 61 (S7) (2005) 176–182.
[33] K. Peng, P. Radivojac, S. Vucetic, A.K. Dunker, Z. Obradovic, Length-dependent prediction of protein intrinsic disorder, BMC Bioinformatics 7 (2006) 1–17.
[34] K. Peng, S. Vucetic, P. Radivojac, C.J. Brown, A.K. Dunker, Z. Obradovic, Optimizing long intrinsic disorder predictors with protein evolutionary information, J. Bioinforma. Comput. Biol. 3 (01) (2005) 35–60.
[35] P. Romero, Z. Obradovic, X. Li, E.C. Garner, C.J. Brown, A.K. Dunker, Sequence complexity of disordered protein, Proteins: Struct., Funct., Bioinf. 42 (1) (2001) 38–48.
[36] B. Xue, R.L. Dunbrack, R.W. Williams, A.K. Dunker, V.N. Uversky, Pondr-fit: a meta-predictor of intrinsically disordered amino acids, Biochim. Biophys. Acta, Proteins Proteomics 1804 (4) (2010) 996–1010.
[37] G.W. Dayhoff, V.N. Uversky, Rapid prediction and analysis of protein intrinsic disorder, Protein Sci. 31 (12) (2022) e4496.
[38] K. Rajagopalan, S.M. Mooney, N. Parekh, R.H. Getzenberg, P. Kulkarni, A majority of the cancer/testis antigens are intrinsically disordered proteins, J. Cell. Biochem. 112 (11) (2011) 3256–3267.
[39] V.N. Uversky, Analyzing idps in interactomes, Methods and Protocols, Intrinsically Disordered Proteins, 2020, pp. 895–945.
[40] V. N. Uversky, J. R. Gillespie, A. L. Fink, Why are "natively unfolded" proteins unstructured under physiologic conditions?, Proteins: Structure, Function, and Bioinformatics 41 (3) (2000) 415–427.
[41] C.J. Oldfield, Y. Cheng, M.S. Cortese, C.J. Brown, V.N. Uversky, A.K. Dunker, Comparing and combining predictors of mostly disordered proteins, Biochemistry 44 (6) (2005) 1989–2000.
[42] A. Mohan, W.J. Sullivan Jr., P. Radivojac, A.K. Dunker, V.N. Uversky, Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes, Mol. BioSyst. 4 (4) (2008) 328–340.
[43] F. Huang, C. Oldfield, J. Meng, W.-L. Hsu, B. Xue, V.N. Uversky, P. Romero, A. K. Dunker, Subclassifying disordered proteins by the ch-cdf plot method, in: Biocomputing 2012, World Scientific, 2012, pp. 128–139.
[44] Z. Chen, P. Zhao, F. Li, A. Leier, T.T. Marquez-Lago, Y. Wang, G.I. Webb, A.I. Smith, R.J. Daly, K.-C. Chou, et al., ifeature: a python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 34 (14) (2018) 2499–2502.
[45] M. Forina, C. Armanino, V. Raggio, Clustering with dendrograms on interpretation variables, Anal. Chim. Acta 454 (1) (2002) 13–19.
[46] A.K. Dunker, J.D. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, C. J. Oldfield, A.M. Campen, C.M. Ratliff, K.W. Hipps, et al., Intrinsically disordered protein, J. Mol. Graph. Model. 19 (1) (2001) 26–59.
[47] R. Williams, Z. Obradovic, V. Mathura, W. Braun, E. Garner, J. Young, S. Takayama, C.J. Brown, A.K. Dunker, The protein non-folding problem: amino acid determinants of intrinsic order and disorder, in: Biocomputing 2001, World Scientific, 2000, pp. 89–100.
[48] P. Radivojac, L.M. Iakoucheva, C.J. Oldfield, Z. Obradovic, V.N. Uversky, A. K. Dunker, Intrinsic disorder and functional proteomics, Biophys. J. 92 (5) (2007) 1439–1456.
[49] S. Barik, The uniqueness of tryptophan in biology: properties, metabolism, interactions and localization in proteins, Int. J. Mol. Sci. 21 (22) (2020) 8776.
[50] P.Y. Chou, G.D. Fasman, Structural and functional role of leucine residues in proteins, J. Mol. Biol. 74 (3) (1973) 263–281.
[51] O. Weiss, M.A. Jiménez-Montaño, H. Herzel, Information content of protein sequences, J. Theor. Biol. 206 (3) (2000) 379–386.
[52] A.L. Berman, E. Kolker, E.N. Trifonov, Underlying order in protein sequence organization, Proc. Natl. Acad. Sci. 91 (9) (1994) 4044–4047.
[53] R.O. Emerson, J.H. Thomas, Adaptive evolution in zinc finger transcription factors, PLoS Genet. 5 (1) (2009) e1000325.
[54] U. Paila, R. Kondam, A. Ranjan, Genome bias influences amino acid choices: analysis of amino acid substitution and re-compilation of substitution matrices exclusive to an at-biased genome, Nucleic Acids Res. 36 (21) (2008) 6664–6675.

[55] A.E. Todd, C.A. Orengo, J.M. Thornton, Evolution of protein function, from a structural perspective, Curr. Opin. Chem. Biol. 3 (5) (1999) 548–556.
[56] V.N. Uversky, A decade and a half of protein intrinsic disorder: biology still waits for physics, Protein Sci. 22 (6) (2013) 693–724.
[57] B. Reva, Y. Antipin, C. Sander, Determinants of protein function revealed by combinatorial entropy optimization, Genome Biol. 8 (11) (2007) 1–15.
[58] A. Jain, H. Gali, D. Kihara, Identification of moonlighting proteins in genomes using text mining techniques, Proteomics 18 (21−22) (2018) 1800083.